

Don Harris (Ed.)

LNAI 10906

Engineering Psychology and Cognitive Ergonomics

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



 Springer

LNAI Series Editors

Randy Goebel

University of Alberta, Edmonton, Canada

Yuzuru Tanaka

Hokkaido University, Sapporo, Japan

Wolfgang Wahlster

DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann

DFKI and Saarland University, Saarbrücken, Germany

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

Don Harris (Ed.)

Engineering Psychology and Cognitive Ergonomics

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

Editor
Don Harris
Coventry University
Coventry
UK

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-91121-2 ISBN 978-3-319-91122-9 (eBook)
<https://doi.org/10.1007/978-3-319-91122-9>

Library of Congress Control Number: 2018942174

LNCS Sublibrary: SL7 – Artificial Intelligence

© Springer International Publishing AG, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

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

Foreword

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

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

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

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

July 2018

Constantine Stephanidis

HCI International 2018 Thematic Areas and Affiliated Conferences

Thematic areas:

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

Affiliated conferences:

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

Conference Proceedings Volumes Full List

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

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

<http://2018.hci.international/proceedings>



Engineering Psychology and Cognitive Ergonomics

Program Board Chair(s): **Don Harris, UK**

- Henning Boje Andersen, Denmark
- Summer L. Brandt, USA
- Oliver Carsten, UK
- Nicklas Dahlstrom, UAE
- Shan Fu, P.R. China
- Wen-Chin Li, UK
- Andreas Luedtke, Germany
- Jan Noyes, UK
- Ling Rothrock, USA
- Axel Schulte, Germany
- Frederic Vanderhaegen, France

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

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



HCI International 2019

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

General Chair

Prof. Constantine Stephanidis

University of Crete and ICS-FORTH

Heraklion, Crete, Greece

E-mail: general_chair@hcii2019.org

<http://2019.hci.international/>



Contents

Mental Workload and Human Error

| | |
|---|-----|
| Design and Evaluation of a Workload-Adaptive Associate System for Cockpit Crews | 3 |
| <i>Yannick Brand and Axel Schulte</i> | |
| The Influence of Culture on Vigilance Performance and Subjective Experience | 19 |
| <i>Qin Gao, Man Wu, and Bin Zhu</i> | |
| The Impact of Metacognitive Monitoring Feedback on Mental Workload and Situational Awareness | 32 |
| <i>Jung Hyup Kim</i> | |
| A Heterarchical Urgency-Based Design Pattern for Human Automation Interaction | 42 |
| <i>Axel Schulte, Diana Donath, Douglas S. Lange, and Robert S. Gutzwiller</i> | |
| A Multidimensional Workload Assessment Method for Power Grid Dispatcher | 55 |
| <i>Bingbing Song, Zhen Wang, Yanyu Lu, Xiaobi Teng, Xinyi Chen, Yi Zhou, Hai Ye, and Shan Fu</i> | |
| Task-Load Evaluation Method for Maintenance Personnel Based on the JACK Simulation. | 69 |
| <i>Ruishan Sun, Yuting Zhang, Zhen Liu, and Kang Li</i> | |
| The Identification of Human Errors in the Power Dispatching Based on the TRACER Method | 80 |
| <i>Xiaobi Teng, Yanyu Lu, Zhen Wang, Bingbing Song, Hai Ye, Yi Zhou, and Shan Fu</i> | |
| Ergonomic Evaluation Study of Occupant Function Allocation for Riot Vehicle Based on Task Load | 90 |
| <i>Qun Wang, Fang Xie, Runing Lin, Xiaoping Jin, and Xue Shi</i> | |
| Effect of Fatigue and Nervousness of Tower Controller on the Control Efficiency | 100 |
| <i>Xingjian Zhang, Peng Bai, Xinglong Wang, and Yifei Zhao</i> | |

Situation Awareness, Training and Team Working

Dynamic Prediction Model of Situation Awareness in Flight Simulation 115
Chuanyan Feng, Xiaoru Wanyan, Shuang Liu, Damin Zhuang, and Xu Wu

The Effect of Thirty-Six Hour Total Sleep Deprivation on Spatial Cognition and Alertness. 127
Wenjuan Feng, Ruishan Sun, and Kai Zhang

Human-Centered Design of Flight Mode Annunciation for Instantaneous Mode Awareness 137
Andreas Horn, Wen-Chin Li, and Graham Braithwaite

Inter-sector Backup Behaviors in Parallel Approach ATC: The Effect of Job Satisfaction. 147
Yazhe Li, Xiaotian E, Han Qiao, Xiangying Zou, Chunhui Lv, Lin Xiong, Xianghong Sun, and Jingyu Zhang

Quantitative Study of Alertness During Continuous Wakefulness Under the Effect of Nervous Activity 158
Kang Li, Ruishan Sun, Jingqiang Li, and Yu-Ting Zhang

Tracking Provenance in Decision Making Between the Human and Autonomy 171
Crisrael Lucero, Braulio Coronado, Eric Gustafson, and Douglas S. Lange

Cyber Officer Profiles and Performance Factors 181
Ricardo G. Lugo and Stefan Sütterlin

Displaced Interactions in Human-Automation Relationships: Transparency over Time. 191
Christopher A. Miller

Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance 204
Mark A. Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen

Crew Resource Management for Automated Teammates (CRM-A) 215
Robert J. Shively, Joel Lachter, Robert Koteskey, and Summer L. Brandt

An Integrated After Action Review (IAAR) Approach: Conducting AARs for Scenario-Based Training Across Multiple and Distinct Skill Areas 230
Lisa Townsend, Joan Johnston, William A. Ross, Laura Milham, Dawn Riddle, and Henry Phillips

How Shared Screen Affected Team Collaboration Task, A Case Study of Ergonomics Experiment on Team Situation Awareness 241
Xu Wu, Chuanyan Feng, Xiaoru Wanyan, Shuang Liu, Lin Ding, Chongchong Miao, Yuhui Wang, and Xueli He

Effect of Different Information Push Mechanism on Driver’s Situation Awareness 250
Bowen Zheng, Xiaoping Jin, Zhenghe Song, Yeqing Pei, and Xuechao Ma

Psychophysiological Measures and Assessment

Mental Workload Estimation from EEG Signals Using Machine Learning Algorithms 265
Baljeet Singh Cheema, Shabnam Samima, Monalisa Sarma, and Debasis Samanta

Psycho-Physiological Evaluation of the Pilot: A Study Conducted with Pilots of the French Air Force 285
Vincent Ferrari, Jean-François Gagnon, Cyril Camachon, and Maëlle Kopf

Variation in Pupil Diameter by Day and Time of Day 296
Shannon R. Flynn, Jacob S. Quartuccio, Ciara Sibley, and Joseph T. Coyne

Computer-Based Neuropsychological Assessment: A Validation of Structured Examination of Executive Functions and Emotion 306
Gilberto Galindo-Aldana, Victoria Meza-Kubo, Gustavo Castillo-Medina, Israel Ledesma-Amaya, Javier Galarza-Del-Angel, Alfredo Padilla-López, and Alberto L. Morán

The Mapping Between Hand Motion States Induced by Arm Operation and Surface Electromyography 317
Tingting Hou, Chen Qian, Yanyu Lu, and Shan Fu

Short Paper: Damage Mechanism and Risk Control on Kid’s Sunglasses 330
Xia Liu, Bisong Liu, Bao Liu, Youyu Xiao, and Yongnan Li

Affective Recognition Using EEG Signal in Human-Robot Interaction. 336
Chen Qian, Tingting Hou, Yanyu Lu, and Shan Fu

Research on Test of Anti-G Suits Airbag Pressure 352
Ding Yi, Zhaowei Zhu, Wang Yandong, Zhang Zhongji, Song Kaiyuan, and Ding Li

Interaction, Cognition and Emotion

The Effects of Risk and Role on Users’ Anticipated Emotions
in Safety-Critical Systems 369
*Yusuf Albayram, Mohammad Maifi Hasan Khan, Theodore Jensen,
Ross Buck, and Emil Coman*

Comparison of Intellectus Statistics and Statistical Package for the Social
Sciences: Differences in User Performance Based on Presentation
of Statistical Data 389
Allen C. Chen, Sabrina Moran, Yuting Sun, and Kim-Phuong L. Vu

Comparative Study of Laptops and Touch-Screen PCs for Searching
on the Web 403
*Nicolas Debue, Cécile van de Leemput, Anish Pradhan,
and Robert Atkinson*

A Pilot Study on Gaze-Based Control of a Virtual Camera
Using 360°-Video Data 419
*Jutta Hild, Edmund Klaus, Jan-Hendrik Hammer, Manuel Martin,
Michael Voit, Elisabeth Peinsipp-Byma, and Jürgen Beyerer*

Efficiency and User Experience of Gaze Interaction
in an Automotive Environment 429
*Benedikt Lux, Daniel Schmidl, Maximilian Eibl, Bastian Hinterleitner,
Patricia Böhm, and Daniel Isemann*

Investigation of Factors Affecting the Usability Evaluation
of an Adaptive Cruise Control System. 445
*Akihiro Maehigashi, Kazuhisa Miwa, Hirofumi Aoki,
and Tatsuya Suzuki*

Accent and Gender Bias in Perceptions of Interactive Voice Systems 457
Sabrina Moran, Ezekiel Skovron, Matthew Nare, and Kim-Phuong L. Vu

Tangible User Interface 471
Elias Shamilov, Nirit Gavish, Hagit Krisher, and Eran Horesh

Population Stereotypes for Color Associations 480
Yuting Sun and Kim-Phuong L. Vu

Presentation of Personal Health Information for Consumers:
An Experimental Comparison of Four Visualization Formats 490
Da Tao, Juan Yuan, Xingda Qu, Tieyan Wang, and Xingyu Chen

Micro and Macro Predictions: Using SGOMS to Predict Phone App Game Playing and Emergency Operations Centre Responses 501
Robert West, Lawrence Ward, Kate Dudzik, Nathan Nagy, and Fraydon Karimi

Natural Interaction in Video Image Investigation and Its Evaluation 520
Yan Zheng and Guozhen Zhao

An Experiment Study on the Cognitive Schema of Trajectory in Dynamic Visualization 533
Xiaozhou Zhou, Chengqi Xue, Congzhe Chen, and Haiyan Wang

Cognition in Aviation and Space

Playbook for UAS: UX of Goal-Oriented Planning and Execution. 545
Jack Gale, John Karasinski, and Steve Hillenius

Augmented Reality in a Remote Tower Environment Based on VS/IR Fusion and Optical Tracking. 558
Maria Hagl, Maik Friedrich, Anne Papenfuss, Norbert Scherer-Negenborn, Jörn Jakobi, Tim Rambau, and Markus Schmidt

Network Re-analysis of Boeing 737 Accident at Kegworth Using Different Potential Crewing Configurations for a Single Pilot Commercial Aircraft 572
Don Harris

Human Performance Assessment of Multiple Remote Tower Operations Simultaneous Take-Off and Landing at Two Airports 583
Peter Kearney, Wen-Chin Li, and Graham Braithwaite

CONTACT: A Human Centered Approach of Multimodal Flight Deck Design and Evaluation. 593
Anne-Claire Large, Cedric Bach, and Guillaume Calvet

A System for Evaluating Pilot Performance Based on Flight Data. 605
Sha Liu, Youxue Zhang, and Jintao Chen

Pilot Performance Assessment in Simulators: Exploring Alternative Assessment Methods 615
Pete McCarthy and Arnar Agnarsson

Now You See It, Now You Don't: A Change Blindness Assessment of Flight Display Complexity and Pilot Performance 637
Claire McDermott Ealding and Alex Stedmon

Experimental Evaluation of a Scalable Mixed-Initiative Planning Associate for Future Military Helicopter Missions 649
Fabian Schmitt and Axel Schulte

Flight Safety: ESL Flight Crew Member Use of Crew Alerting
and Information Systems. 664
Dujuan Sevillian

The Preliminary Application of Observer XT(12.0)
in a Pilot-Behavior Study. 686
Ruishan Sun, Guanchao Zhang, and Zhibo Yuan

Tablet-Based Information System for Commercial Aircraft: Onboard
Context-Sensitive Information System (OCSIS). 701
Wei Tan and Guy A. Boy

Modeling and Simulating Astronaut’s Performance
in a Three-Level Architecture. 713
*Chunhui Wang, Shanguang Chen, Yuqing Liu, Dongmei Wang,
Shoupeng Huang, and Yu Tian*

Risk Cognition Variables and Flight Exceedance Behaviors
of Airline Transport Pilots 725
Lei Wang, Jingyi Zhang, Hui Sun, and Yong Ren

Author Index 739

Mental Workload and Human Error



Design and Evaluation of a Workload-Adaptive Associate System for Cockpit Crews

Yannick Brand^(✉) and Axel Schulte

Institute of Flight Systems, University of the Bundeswehr Munich,
Neubiberg, Germany


{y.brand, axel.schulte}@unibw.de

Abstract. This article describes and validates a concept of a workload-adaptive associate system for military helicopter crews. We use adaptive automation to support helicopter pilots during Manned-Unmanned Teaming missions, where the crew of a manned helicopter operates several unmanned aerial vehicles from the cockpit. We introduce a cognitive agent, which behaves like an additional, artificial crew member. It dynamically adjusts its level of assistance by choosing different workload-adapted strategies of assistive intervention depending on free mental resources of the crew. To evaluate the prototype, we conducted an extensive pilot-in-the-loop campaign and analyze situations of “near misses”, where the associate system corrects human erroneous behavior.

Keywords: Adaptive automation · Associate system · Cockpit automation
Workload-adaptive · Human factors · Human-agent teaming

1 Introduction

At the Institute of Flight Systems, the concept of Manned-Unmanned Teaming (MUM-T) is a well-established approach to improve future military aviation. Guiding multiple reconnaissance Unmanned Aerial Vehicle (UAV) from the cockpit yields many advantages including information gain, flexibility and safety. But the additional tasks of operating the UAV and process the information of their sensors increase the workload of the cockpit crew. If the demand exceeds the available mental resources (i.e. workload peak), e.g. in time-critical multi-tasking situations, it results in performance decrements and human errors [1]. In the aviation domain, automation has always been a solution to reduce workload. But shifting more tasks from the human to (conventional) automation brings other problems: Loss of situation awareness, workload peaks due to high demanding cognitive tasks, which remain with the human, and “automation induced workload peaks” due to clumsy automation [2].

To counteract this, Onken and Schulte introduce **two modes of cognitive automation** [3, 4]. This automation is able to handle and support cognitive tasks, to avoid exceeding mental resources on the one hand and too little or wrong human involvement on the other hand. One mode is, to establish a cognitive agent (, see Fig. 1) onboard of each UAV which facilitates a task-based (instead of parameter-based)

UAV guidance [5]. The agent is able to understand high-level commands, formulated as tasks, and reports the results back on the same task-based level. Several studies show, that these cognitive agents enable the crew of a manned helicopter to guide multiple UAV from the cockpit while performing complex mission scenarios [6].

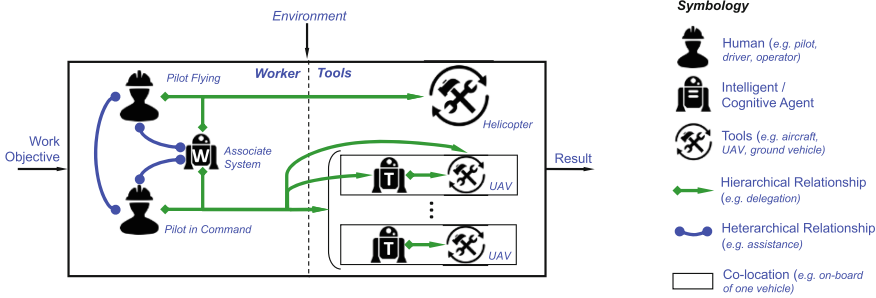


Fig. 1. Work system of the manned-unmanned teaming configuration in the human-agent teaming notation from [4].

As second mode, we introduce a cognitive agent (see Fig. 1) as associate system to support the crew adapted to their mental state. The goal is, to prevent workload peaks and human errors while keeping the crew “in the loop” to prevent out-of-the-loop problems [7], complacency and automation bias [8]. Therefore, the associate system takes the role of an artificial, restrained-behaving co-pilot. Onken and Schulte describe a guideline for this restrained behavior as a set of rules [3]. The basic rule is, to let the crew do their job as long as possible without the associate system intervening. However,

1. if the attention of the human is not on the objectively most important task, guide the human’s attention to that task;
2. if, nevertheless, the human is overtaxed, transform the task situation into one, the human can handle again. And only
3. if the human is in principle not able to perform the task and the cost of a failure would be too high, adopt the task as a last resort.

These escalating behavior rules link the type and amount of assistance to the attention and mental state of the crew. Therefore, the adaptive associate system needs a context-rich representation of mental workload [9]. It is not sufficient to know if the workload is high, but in addition it is necessary to know the reason, i.e. the causal task situation, which leads to high workload. Only the knowledge about the task situation enables the agent to support the crew task-based, also for cognitive tasks.

2 Task-Based Operationalization of the Mental State as Precursor of Adaptive Assistance

As basis for the workload-adaptive associate system we introduce a context-rich definition of mental workload, as described in [9]. It includes

- the currently pursued work objective and the resulting tasks, which are necessary to achieve this objective, i.e. **plan**;
- the set of tasks, which the operator currently executes, i.e. **activity**;
- the **demand on mental resources**, which is necessary to execute the activity and
- **behavior patterns**, which the operator typically shows during the task execution and variations from these patterns [10].

2.1 Task Model

We operationalize this definition of mental workload with tasks. Therefore, we developed a hierarchical task model which contains the domain knowledge of our application, i.e. all tasks which can occur during a MUM-T helicopter mission [11]. This machine-readable task model enables the associate system

- to have sufficient knowledge of the domain, similar to the knowledge of the human crew. This includes the demand on mental resources, which is necessary to execute a specific task, constraints for tasks and relations between tasks;
- to communicate with the crew in a natural manner, since humans communicate very efficient by using tasks as expression for very complex situations;
- to communicate in the same efficient way between different modules of the associate system;
- to know different variants of task sharing between the crew and the automation. According to [1], working memory load can be reduced by automating tasks. That is, increasing the level of automation, means to decrease the involvement and therefore the taskload of the human operator. The **Levels of Automation (LoA)** in our task model define, how the associate system can reduce the crews taskload by automating a specific abstract task higher (like the mission planning task “PilotPlanMission” in Fig. 2).

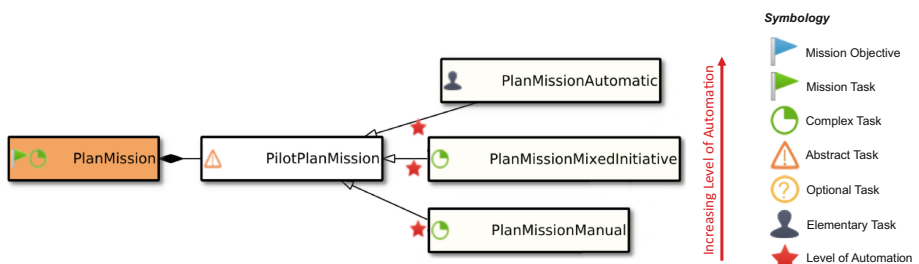


Fig. 2. Excerpt of the task model, which shows the abstract task “PilotPlanMission” and its three different Levels of Automation, marked with red stars. As higher the LoA is, as less the human involvement and therefore taskload of the crew. The image does not show the subtasks of lower levels of the task model.

2.2 Plan

Based on the tasks of the task model, the plan is a dynamically generated sequence of mission tasks (▶, see example mission plan in Fig. 3), which fulfills the mission objective (▶, see Fig. 3). In our application, the plan depends on the objective of the MUM-T helicopter mission and constraints like terrain, air spaces and others. To plan and schedule this sequence, we use a mixed-initiative approach, where the pilot is in charge to plan the mission for reasons of plan situation awareness, transparency and trust. The mixed-initiative mission planner (MIP) supports the planning process by intervening on own initiative in case of threat avoiding, missing tasks or optimization [12]. The MIP knows different levels of automation (the three levels depicted in Fig. 2), which the associate system can use to simplify planning related problems.

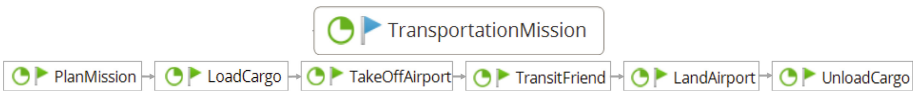


Fig. 3. Example of a transport helicopter mission. The order Transportation Mission can be fulfilled by the complex mission tasks on the bottom.

2.3 Activity

Beside the plan, the activity is a key element in our definition of the crew’s mental state. Therefore, our activity determination observes the crew to identify the current activity on-line [13]. It uses different measurement sensors – gaze tracking and manual interaction tracking as well as speech recognition – and the knowledge about observable evidences from the task model. It combines all simultaneously observed evidences by using an evidential reasoning approach derived from Dempster-Shafer theory to assign each evidence to a task from the task model [13, 14]. The activity is the set of elementary tasks (▶, see Fig. 2), which the crew executes at a given moment. We assume, that if the percentage, which supports the execution of a specific task (blue belief value in Fig. 4) is greater than 0.5, the task is part of the activity. Therefore, in the example in Fig. 4, CommunicateIntern and FlyTransitFriendManual is the current activity of the pilot.

Right Pilot

Detected Activity: CommunicateIntern and FlyTransitFriendManual

| | Task | Filtered | Belief | Doubt | Ignorance |
|---|------------------------|----------|--------|-------|-----------|
| 1 | CommunicateIntern | | 0.97 | 0.03 | 0.00 |
| 2 | FlyTransitFriendManual | | 0.95 | 0.05 | 0.00 |
| 3 | CheckRadio | | 0.20 | 0.00 | 0.80 |
| 4 | CommunicateATC | | 0.00 | 1.00 | 0.00 |

Fig. 4. Screenshot of an example result of the activity determination. Here, the pilot is flying a manual transit flight and communicating via the intercom at the same time (taken from [13]). The belief value of CheckRadio and CommunicateATC is too low to be part of the activity.

2.4 Demand on Mental Resources

To be able to estimate the demand on mental resources, which is necessary to execute the entire activity, the task model stores the demand on mental resources for each single task as demand vector [15]. After identifying the activity, our resource assessment combines all related demand vectors by using the conflict matrix of Wickens' Multi-Resource Theory to estimate an overall workload value [11, 15, 16] (Fig. 5). Due to the eight separately stored components of the demand vector for the different resources (visual-spatial, visual-verbal, auditory-spatial, auditory-verbal, cognitive-spatial, cognitive-verbal, response-manual and response-vocal), our method provides information, **which resource of the pilot** leads to a workload peak. And, as mentioned before, from the activity determination, the associate system knows **which task situation** leads to that workload peak.

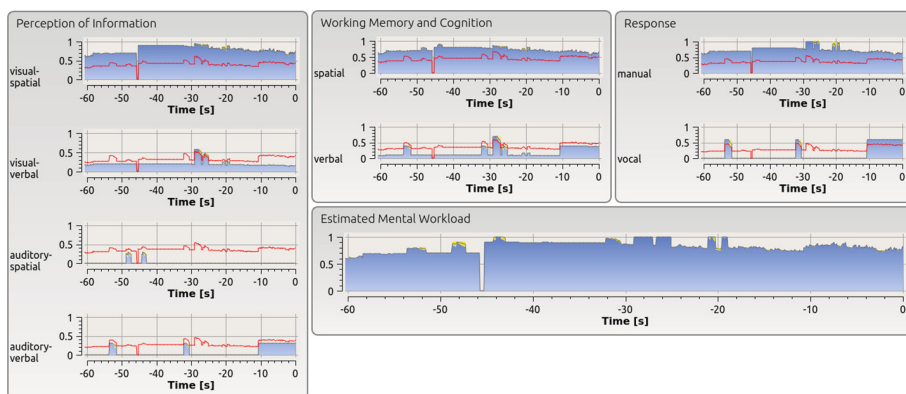


Fig. 5. Screenshot of the resource assessment, which shows the mental resources usage on the three stages – perception, cognition and response – as well as the total mental workload estimation over the last 60 s for the activity shown in Fig. 4.

3 Concept and Implementation of Adaptive Assistance

The context-rich representation of mental workload is the basis for adaptive interventions, which support the crew adapted to their mental state for specific task situations.

Figure 6 depicts the functional architecture of the associate system. The process of inferring adaptive interventions consists of three phases. Before the associate system is able to identify the trigger for assistance – neglected task, (predicted) workload peaks and critical events – it preprocesses the mental state (Phase ① in Fig. 6). Thereby, it uses the plan to project the mental state into the future, to

- know which elementary tasks are necessary to fulfill the plan and
- identify situations of high workload in the future, which may occur during the plan execution.

In the second step, the associate system identifies trigger for adaptive assistance (“critical states”) and plans the intervention (Phase ② in Fig. 6). To implement the adaptive assistance, the associate system uses the human-dialog interface, which is part of the human-machine interface. In addition, the associate system uses other (cognitive) automation (e.g. the mixed-initiative planner or the adaptive crew sensor interaction) to simplify task situations or adopt tasks via the automation dialog interface (Phase ③ in Fig. 6). For detailed information about the process of identifying trigger, the decision process and stages of intervention see [17].

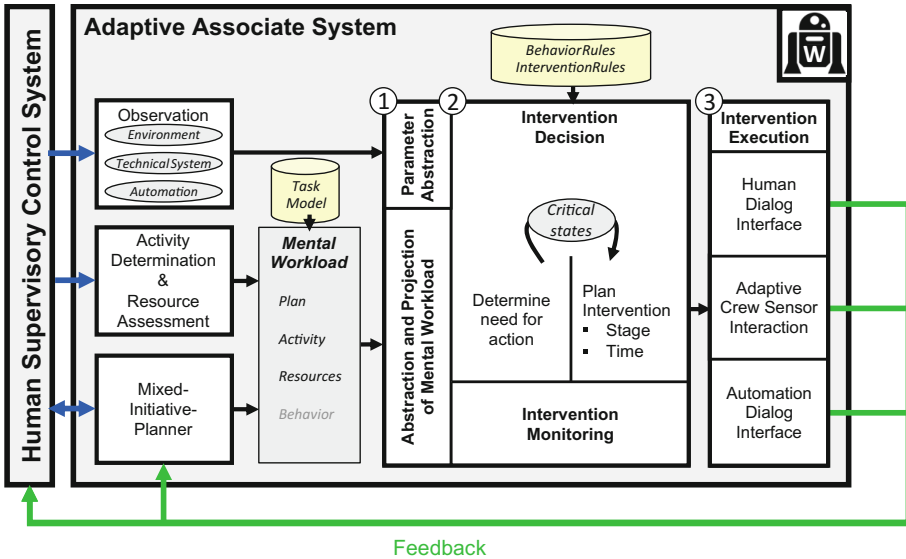



Fig. 6. Functional architecture of the associate system: “This is inside the little  in Fig. 1.” The process of inferring adaptive interventions has three phases: Mental state preprocessing (1), intervention decision (2) and intervention execution (3).

3.1 Phase 1: Identifying Trigger for Adaptive Assistance

For predicting future mental states and identifying resource conflicting task situations as well as neglected tasks, the associate system refines the plan – which is part of the “input” mental workload (see grey box “Mental Workload” in Fig. 6) – to a detailed plan on elementary task level (which is the lowest level of the hierarchical task model). It uses the task relationships from the task model to identify all elementary tasks, which belong to a specific mission task. Figure 7 shows the mission task EnterHOA, which is the task of entering the helicopter mission area, and its subtasks after the refinement process.

Because the activity determination expresses the activity on the same level of elementary tasks, the associate system is able to match the activity with the detailed plan and check the completed tasks (green task-boxes in Fig. 7). All tasks which are

planned, but not executed timely are neglected task and trigger for adaptive assistance (red task-boxes in Fig. 7).

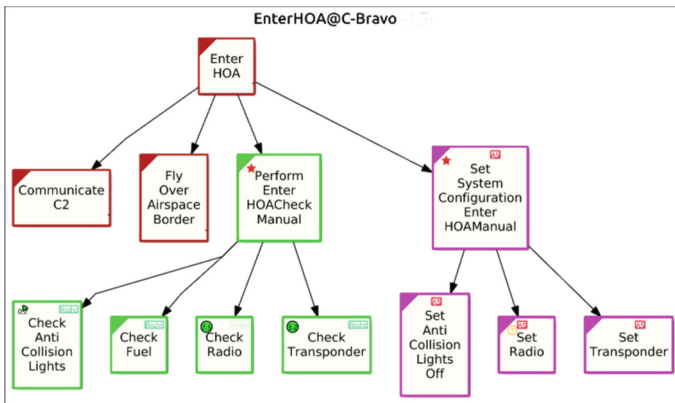


Fig. 7. Excerpt of the task model displaying the hierarchical sub-task relationships of the mission task EnterHOA. The green boxes indicate completed, the red not completed tasks. If the associate system adopts a task, the box becomes magenta. (Color figure online)

In a second step, the associate system schedules all elementary tasks. Therefore, it uses the execution time and task constraints from the task model. This results in a task timeline, as Fig. 8 shows. The associate system simulates the execution of all future task situations and estimate the needed demand on mental resources (red graph curve in Fig. 8). A situation in which the crew has to perform many tasks in parallel and therefore their demand on mental resources exceeds a threshold is a workload peak (red marked period in Fig. 8) and trigger for adaptive assistance.

The third trigger for adaptive assistance are critical events, like changes of the tactical situation, failures of the helicopter systems or threatened UAV. The associate system cannot predict such events. However, if they occur, it supports the crew by guiding the attention, offering a possible solution, or directly implementing a solution.

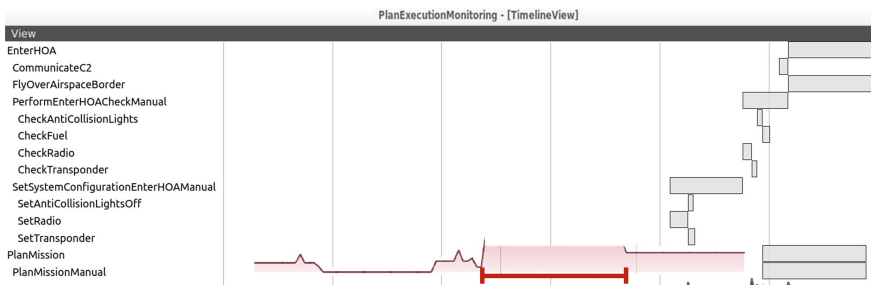


Fig. 8. Timeline of the mission task EnterHOA and its subtasks (the first 11 grey task-boxes) and the mission task PlanMission and its subtasks. Due to many parallel tasks, the associate system predicts a workload peak (red marked time period). (Color figure online)

The following section explains which stage of intervention (which is the “amount of help”) is appropriate.

3.2 Phase 2: Deriving Adaptive Interventions

The process of deriving adaptive interventions aims to find the appropriate level of assistance for a given problem by taking the workload, attention and criticality of the problem into account. Basis for this is the set of rules, mentioned in the introduction. To implement the restrained behavior, the associate system follows the decision process in Fig. 9: First, it traces back each trigger (grey boxes on the left side) to its causal task(s). After comparing this task with the current activity, the associate system infers if the crew works on the solution (Question A in Fig. 9). If the crew is currently not solving the problem, the associate system simulates, whether the crew can handle the current task situation including all other tasks, which are necessary to solve the problem (Question B in Fig. 9). Therefore, it combines the demand vectors of the hypothetical activity using the same method as for the resource assessment, described in Sect. 2.4. If the crew can handle the entire task situation, the associate system guides the attention to the problem (Stage 1 on the right side in Fig. 9), but if not, more assistance is necessary. If the crew is already overtaxed, or any additional task would overtax them, the associate system simplifies the task situation by adapting the human-machine interface (Stage 2a in Fig. 9) or changing the level of automation (Stage 2b in Fig. 9). Only if the problem poses an extreme or high risk (Question C in Fig. 9) and the crew cannot handle it, the associate system is allowed to adopt the problem-solving task(s) (Stage 3 in Fig. 9). We implemented this decision process with the cognitive framework Soar [18]. For detailed information about the decision process and its implementation see [17].

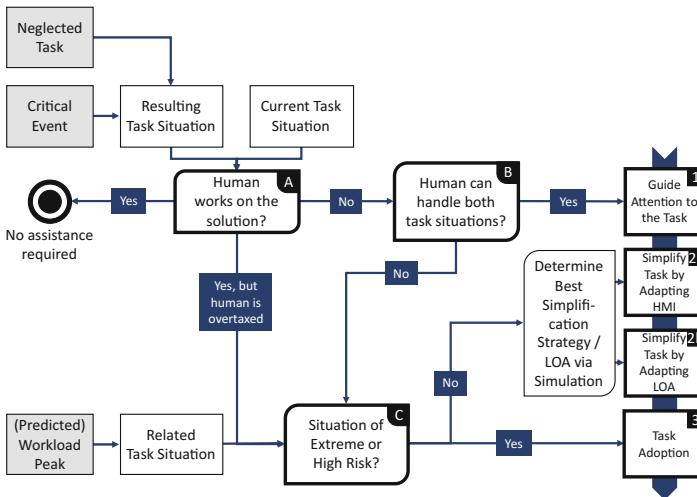


Fig. 9. Decision process of determining the appropriate stage of intervention (bold boxes on the right) based on the related trigger (grey boxes on the left) and the mental state of the crew.

3.3 Phase 3: Stages of Adaptive Interventions

After identifying the appropriate stage of intervention, the associate system implements the different stages as follows. For guiding the attention to a problem, it uses the human-dialog interface. Therefore, it overlays a dialog box on the multi-function displays (MFD) of the cockpit (see left image in Fig. 10) and highlights all related objects on the MFD. In addition, it can simplify the task situation by adding a “short-cut button” to the dialog box, which implements the proposed solution with one button click (second image from left in Fig. 10). This “short-cut button” is also available via the helicopter control stick. Another possibility to simplify the task situation is to change the level of automation, e.g. to increase the level of automation of the mixed-initiative planner to simplify a planning related problem. For critical situations, the associate system can adopt task(s) like a forgotten landing check (right lower image in Fig. 10), if the human is not capable to handle it. The MFD displays the information on a “history list of adopted tasks” (right upper image in Fig. 10). In addition, a speech synthesizer announces the adopted task (e.g. Set System Configuration Landing) as a human team mate would do in this situation.

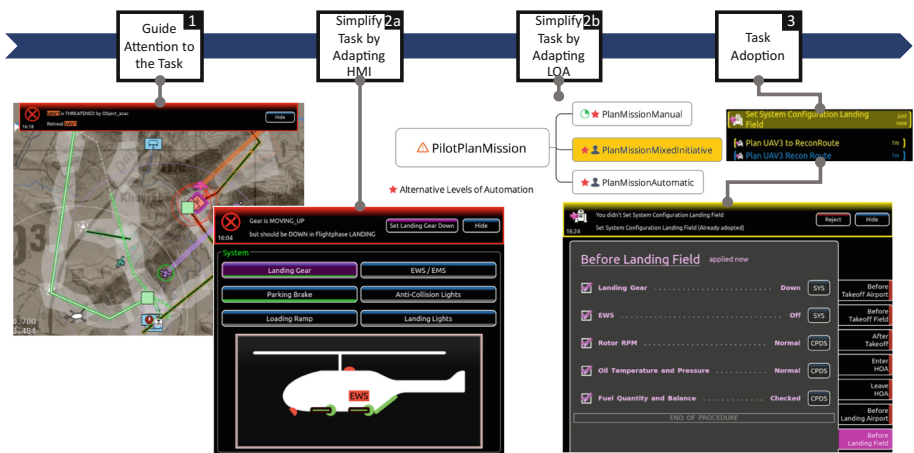


Fig. 10. Stages of intervention, escalating from left to right: Attention Guiding (1), Task Simplification (2a and 2b) and Task Adoption (3)

4 Evaluation

To evaluate the associate system, we conducted a human-in-the-loop experimental campaign with trained military helicopter pilots in our helicopter mission simulator at the Institute of Flight Systems (see Fig. 11). The purpose of the experiment was, to investigate the interventions of the adaptive associate system during realistic and very complex MUM-T mission scenarios. Therefore, in Sect. 4.2 we analyze two (of many occurred) situations of “near misses”, where the associate system corrected human erroneous behavior during the missions. In addition, in Sect. 4.3 we present and analyze the overall system rating given by the pilots.



Fig. 11. Helicopter mission simulator at the Institute of Flight Systems, which we use to implement and evaluate the associate system for MUM-T missions.

4.1 Experimental Design

Our aim at the Institute of Flight Systems is, to evaluate new concepts by testing them as highly integrated systems in very realistic and immersive scenarios. Therefore, the results typically have a very good external validity. But the high complexity of such systems requires domain experts as test subjects (i.e. experienced military helicopter pilots) and moreover an extensive training. We trained our pilots for about two days before conducting experimental trials for another three days. The study population comprises an overall number of seven participants (Age: M 50.4 SD 9.2, Flight Hours: M 3933 SD 1807) grouped into four crews. Every crew consists of a pilot in command (PIC), who leads the mission and guides three UAV for reconnaissance purposes, and a pilot flying (PF), who is responsible to fly the helicopter, communicate and manage the systems of the helicopter. One participant was part of two crews (one time as PIC and another time as PF). Since the associate system supports both crewmember in very different ways, this participant could rate the associate system two times.

The crews performed six different transport helicopter missions (Mission duration in minutes: M 46 SD 10), with mission elements like troops transport, Medical Evacuation (MedEvac) and Combat Search and Rescue (CSAR). The missions are very complex, since all contain many events like suddenly occurring enemies and mission goal changes. These compacted, challenging missions provoke human errors. The intention behind that: In normal workload conditions, highly trained pilots perform good and human errors are very rare. However, to evaluate the associate system, the

deliberate provocation of errors is necessary. Therefore, our missions are designed to represent the most stressful parts of helicopter missions. The pilots confirmed that the missions are very stressful, but nevertheless they rated the scenarios as realistic (Scenario is realistic: M 5.6 SD 0.7 on a 7-point Likert scale).

We did not vary and compare configurations of the system, but analyze, when, why and how the associate system intervenes to help the crew and how the crew rates these interventions.

4.2 Adaptive Interventions

One intervention affects the pilot in command (PIC), who is responsible for planning the mission and operating the UAV to ensure a reconnoitered flight route for the helicopter. In this situation, the PIC forgets to assign the route reconnaissance task for the next flight leg of the helicopter route to a UAV (highlighted in magenta in Figs. 12 and 13). In addition, he is involved in an ongoing route reconnaissance task of another UAV (see green frame in Fig. 12, which shows the tasks of the detailed plan for the next minutes). The two parallel tasks PlanMission and ReconRoute, starting at the red line in Fig. 12, lead to a predicted workload peak for the immediate future (red marked area, “now” is at the left border of the red marked area). To relieve the task situation for the PIC, the associate system instructs the mixed-initiative mission planner to increase the level of automation (remember the planning automation levels ★ in Fig. 2) and to propose the next relevant tasks on its own initiative. The planner proposes the forgotten route reconnaissance task (see dialog box in Fig. 13). Due to the urgency and criticality of this task, the associate system automatically accepts the planner proposal. The pilot rated this workload-adaptive intervention as very helpful and appropriate for this risky and time-critical situation.

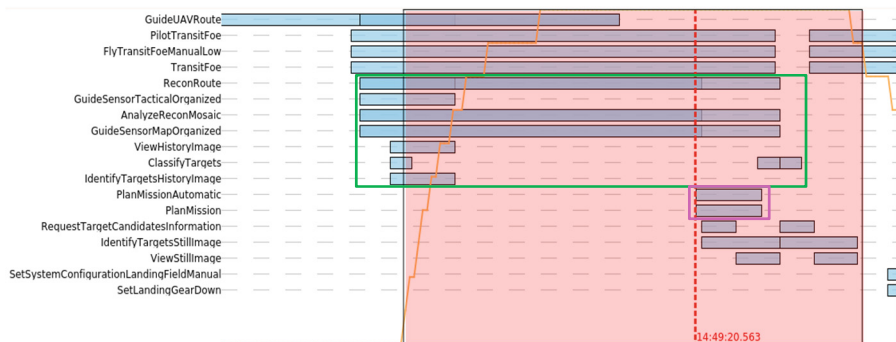


Fig. 12. Prediction of the task situation in the near future (blue blocks are the tasks) and estimation of the future workload (orange solid line) with predicted workload peak (red marked area) and missing UAV task (PlanMission task in magenta frame). (Color figure online)

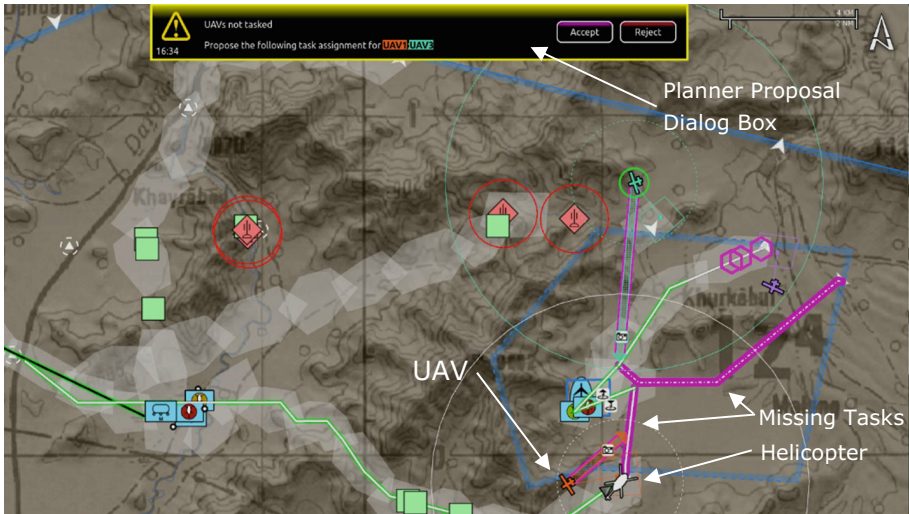


Fig. 13. Screenshot of the tactical map display with the planner proposal dialog box and the missing UAV reconnaissance tasks (highlighted in magenta) which are directly in front of the helicopter symbol, and therefore very urgent. (Color figure online)

One situation in the field of system management affects the pilot flying. Because of a mode confusion during the After-Takeoff checklist, the pilot switches on the landing lights erroneously. Within a possibly threatened area, this is an avoidable safety risk. Therefore, the associate system guides the attention to this wrong configured system state (left blue dashed line in Fig. 14). Because the activity determination enables the agent to recognize, that the pilot flying is doing a low-level flight and has therefore less free resources for doing an additional manual task (see estimated overall workload, represented by the green solid graph in Fig. 14), it decides to simplify the task situation by offering the “short-cut” button. Thereby, the pilot can accept the help with a button on his cyclic stick. The pilot accepts the help via this button (brown dashed line, see Fig. 14) and the associate system switches the lights off (right blue dashed line). As soon as the associate system solves the problem (green dashed line), it notifies the pilot by announcing the related task from the task model (i.e. SetLandingLightsOff). The speech announcement is also visible as task block “ListenAssistanceSpeech” in Fig. 14. The pilot flying rated this intervention as very supporting and helpful. In addition, he stated that the explaining text of the intervention (see Fig. 15) makes it easy to understand the problem and how this problem could arise.

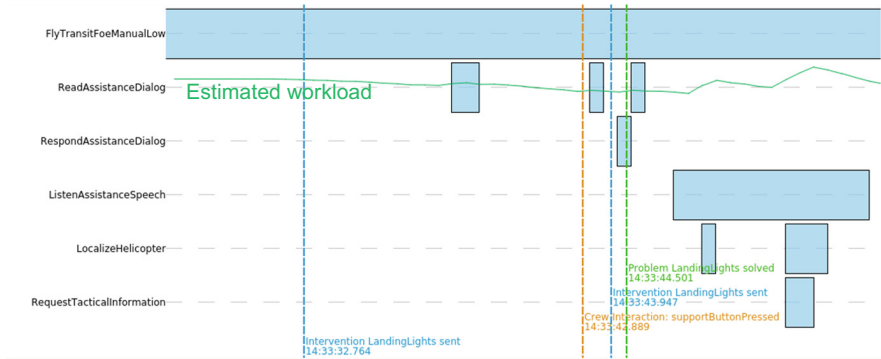


Fig. 14. Task situation (blue blocks) of the pilot flying and overall workload estimation (green solid line reaching from “no workload” at the lower border of the image to “overload” at the upper border of the image) during the intervention (blue dashed line) regarding the landing lights. (Color figure online)

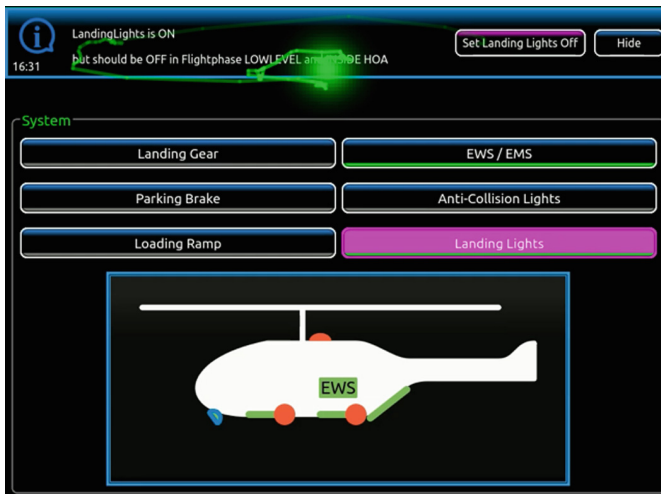


Fig. 15. Screenshot of the system management page of the helicopter simulator, with the green gaze point of the PF (including the local gaze measure accuracy as Gaussian distribution), who is reading the explaining text in the dialog box of the intervention. (Color figure online)

These situations illustrate, that the associate system adaptively supports the crew by taking the (projected) mental state and the criticality into account, and that the pilots feel appropriately supported. Beside the investigation of situations of near misses, the pilots assess the overall performance and behavior of the associate system.

4.3 Results

Most of the pilots rate the supporting interventions of the associate system as expedient and helpful (see Fig. 16). The pilots state, that the interventions are justified and the system reacts correctly in dangerous situations. One strength of the associate system is, that it keeps the overview of the whole situation, if the human is focusing on one task and guide the attention to the most urgent task, e.g. new threats, if necessary. Another major benefit is that the interventions save time by simplifying the task situation, which is very valuable during complex missions. These statements, regarding the attention guiding and task simplification, support the rating, that the stage of intervention is mostly situation-adapted and appropriate (see Fig. 16).

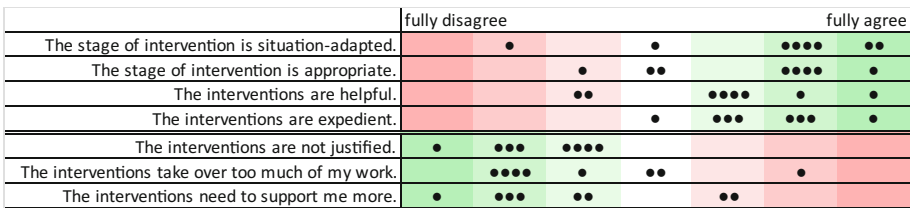


Fig. 16. Overall rating of the interventions on a 7-point Likert scale (n = 8). The dots are the ratings of the pilots.

Although the system knows the current activity, there are rare cases, where one of the pilots and the associate system start working on the same task simultaneously. This is confusing and the pilot needs to spend time to understand this unnecessary intervention. Therefore, the associate system tries to prevent such cases, but due the delay between action planning and action implementation – which both, human and associate system, have – it was not possible to preclude these cases completely. A better situation awareness of the upcoming tasks and the human-agent task sharing, i.e. by pre-announcing of the task sharing, could solve this issue. [19] proposes a possible solution for this kind of pre-announcement.

Most of the pilots appraise the behavior of the associate system as transparent and comprehensible (see Fig. 17). Due to the fact that the associate system intervenes workload-adaptive and there is no pre-announcement of interventions, the predictability is lower. The pilots think, that more training may increase the predictability, but workload-adaptive systems are low predictable at all. As mentioned above, pre-announcing the next tasks and possible interventions increases the situation awareness of the crew [19].

The pilots state, that the associate system is neither hasty nor restrained, but in this case the neutral position is a positive rating, because they state, that they interpret “restrained” as negative like “too restrained”.

| | | | | | | | | |
|------------------|--|----|----|------|------|--------|----|----------------|
| opaque | | • | • | | • | ••••• | | transparent |
| incomprehensible | | | •• | | | •••• | •• | comprehensible |
| unpredictable | | | • | • | •• | •••• | | predictable |
| unreliable | | | • | • | •• | •• | •• | reliable |
| dubious | | | | •• | | •••••• | | trustworthy |
| stupid | | | | •• | •• | •••• | | clever |
| unfamiliar | | | • | • | •••• | •• | | familiar |
| hasty | | •• | • | •••• | • | | | restrained |

Fig. 17. Overall rating of the behavior of the associate system on a 7-point Likert scale ($n = 8$). The dots are the ratings of the pilots.

5 Conclusion

We presented a concept of a workload-adaptive associate system, which supports the crew of a manned helicopter during military MUM-T helicopter missions. The process of inferring workload-adaptive interventions relies on a context-rich operationalization of mental workload based on tasks. With this definition of mental workload, the associate system is able to project future task situations to identify all pilot tasks which are necessary to reach the mission goal. In addition, it predicts situations of high workload in the future and eases them proactively by using higher levels of automation. After identifying trigger for supporting interventions, the associate system supports the crew by guiding the attention, offering a possible solution or directly implementing a solution. Thereby, it behaves restrainedly and helps only if necessary and as less as possible to keep the human in the loop and to prevent typical pitfalls of highly automated systems, like out-of-the-loop problems, complacency and automation bias.

Our pilot-in-the-loop experiments show, that the concept of supporting cognitive tasks workload-adaptively recovers human errors like neglected tasks and relieves time critical task situations before they occur. The pilots rate the interventions of the associate system as helpful and expedient. However, future improvements of the system should address the transparency e.g. by pre-announcing the human-agent task sharing.

We implemented this concept in the domain of military aviation. But the concept is not limited to this domain. It is transferable to other domains, where a human operator collaborates with highly automated systems, e.g. civil aviation, highly automated driving or power plant management.

References

1. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. *Ergonomics* **58**, 1–17 (2015)
2. Wiener, E.L.: Human Factors of Advanced Technology (Glass Cockpit) Transport Aircraft. NASA CR 177528, Ames Research Center, Moffett Field (1989)
3. Onken, R., Schulte, A.: System-Ergonomic Design of Cognitive Automation: Dual-Mode Cognitive Design of Vehicle Guidance and Control Work Systems. *SCI*, vol. 235. Springer, Heidelberg (2010). <https://doi.org/10.1007/978-3-642-03135-9>

4. Schulte, A., Donath, D., Lange, D.S.: Design patterns for human-cognitive agent teaming. In: Harris, D. (ed.) EPCE 2016. LNCS (LNAI), vol. 9736, pp. 231–243. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_24
5. Rudnick, G., Schulte, A.: Implementation of a responsive human automation interaction concept for task-based-guidance systems. In: Harris, D. (ed.) EPCE 2017. LNCS (LNAI), vol. 10275, pp. 394–405. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58472-0_30
6. Uhrmann, J., Schulte, A.: Concept, design and evaluation of cognitive task-based UAV guidance. *Int. J. Adv. Intell. Syst.* **5**, 145–158 (2012)
7. Endsley, M.R., Kiris, E.O.: The out-of-the-loop performance problem and level of control in automation. *Hum. Factors* **37**, 381–394 (1995)
8. Parasuraman, R., Manzey, D.H.: Complacency and bias in human use of automation: an attentional integration. *Hum. Factors* **52**, 381–410 (2010)
9. Schulte, A., Donath, D., Honecker, F.: Human-system interaction analysis for military pilot activity and mental workload determination. In: Proceedings of 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015, pp. 1375–1380 (2015)
10. Donath, D., Schulte, A.: Behavior based task and high workload determination of pilots guiding multiple UAVs. *Procedia Manuf.* **3**, 990–997 (2015)
11. Honecker, F., Brand, Y., Schulte, A.: A task-centered approach for workload-adaptive pilot associate systems. In: Proceedings of the 32nd Conference of the European Association for Aviation Psychology – Thinking High AND Low: Cognition and Decision Making in Aviation, Cascais (2016)
12. Schmitt, F., Roth, G., Schulte, A.: Design and evaluation of a mixed-initiative planner for multi-vehicle missions. In: Harris, D. (ed.) EPCE 2017. LNCS (LNAI), vol. 10276, pp. 375–392. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58475-1_28
13. Honecker, F., Schulte, A.: Automated online determination of pilot activity under uncertainty by using evidential reasoning. In: Harris, D. (ed.) EPCE 2017. LNCS (LNAI), vol. 10276, pp. 231–250. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58475-1_18
14. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
15. Wickens, C.D.: Multiple resources and performance prediction. *Theor. Issues Ergon. Sci.* **3**, 159–177 (2002)
16. Maiwald, F., Schulte, A.: Enhancing military helicopter pilot assistant system through resource adaptive dialogue management. In: Vidulich, M.A., Tsang, P.S., Flach, J.M. (eds.) *Advances in Aviation Psychology*. Ashgate Studies in Human Factors and Flight Operations (2014)
17. Brand, Y., Schulte, A.: Model-based prediction of workload for adaptive associate systems. In: Proceedings of 2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017, pp. 1722–1727 (2017)
18. John, E.: *Laird: The Soar Cognitive Architecture*. The MIT Press, Cambridge (2012)
19. Brand, Y., Ebersoldt, M., Barber, D., Chen, J.Y.C., Schulte, A.: Design and experimental validation of transparent behavior for a workload-adaptive cognitive agent. In: Karwowski, W., Ahram, T. (eds.) *IHSI 2018*. AISC, vol. 722, pp. 173–179. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73888-8_28



The Influence of Culture on Vigilance Performance and Subjective Experience

Qin Gao^(✉), Man Wu, and Bin Zhu

Department of Industrial Engineering, Tsinghua University,
Beijing, People's Republic of China
gaoqin@tsinghua.edu.cn

Abstract. This study conducted a 23-min Continuous Performance Test (CPT) to investigate the impact of cultural background (Chinese and Americans) on vigilance performance and subjective experience. The performance included correct detections, false alarm rate, and reaction time. The subjective experience included perceived workload using NASA-TLX scales and perceived fatigue. Twenty Chinese and twenty Americans participated in the experiment. This study found that American participants performed significantly better than Chinese participants in correct detections and reaction time. Chinese felt more workload and fatigue than Americans. In terms of dimensions, Chinese participants felt higher levels of physical demand, physical fatigue, boredom, and paid more efforts than American participants. The result suggested that Chinese tend to allocate attention in a broad region, whereas Americans are inclined to focus attention on focal objects.

Keywords: Vigilance task · Culture · Performance · Workload

1 Introduction

Vigilance has become increasingly salient as a critical human factors issue because of the increasing automation of technology (Warm et al. 2008). Vigilance tasks require the observer to keep their focused attention on information displays for prolonged periods of time and detect the appearance of predetermined critical signals, which often occur infrequently (Dillard et al. 2014; Funke et al. 2016; Helton and Russell 2011). The recent advancement of automation technology makes vigilance an even more salient issue. The role of workers in an automated system shifts from manual controllers to system supervisors, who need to monitor the system and act only in instances of problems or system failures (Sheridan 1992). Moreover, automation has been introduced into many safety-critical domains such as military surveillance (Dillard et al. 2014; Helton and Russell 2011), aircraft and air traffic control (Langan-Fox et al. 2009), vehicle operation (Young and Stanton 2007), nuclear power plant and process control (Baker et al. 1994), and health care (Drews 2013; Sheridan 2002). Therefore, suboptimal performance in vigilance tasks can lead to severe accidents and consequences, as reported by several studies (Hawley et al. 2006; Molloy and Parasuraman 1996; Warm et al. 2008).

Vigilance decrement is a major finding of prior vigilance research. It has been observed in a variety of industrial, transportation, security, and medical settings (Parasuraman 1986; Warm et al. 2008). Vigilance decrement refers to loss of detection of critical stimuli during a prolonged, continuous work shift (Kamzanova et al. 2014). In laboratory studies, vigilance decrement is commonly observed within the first 20 to 30 min of performance (Molloy and Parasuraman 1996) and on some tasks, over even shorter durations (Gillberg and Åkerstedt 1998; Temple et al. 2000). The resource model theory, proposed by Parasuraman and Davies (1977), posits that the need to make continuous signal-noise discriminations under great uncertainty without rest depletes information-processing resources that cannot be replenished during task performance, leading to a decline in performance efficiency. From the perspective of attention, mindlessness theory asserted that the repetitive and monotonous nature of vigilance tasks lead to a lack of attentional focus on the task (Manly et al. 1999; Robertson et al. 1997). During sustained attention, the attention often drifts to some thoughts unrelated to task over time (Smallwood et al. 2004).

Due to vigilance decrement and the importance of vigilance performance for system safety, many scholar efforts have been devoted to study factors that affect vigilance performance. Some factors are related to features of stimuli, such as frequency of signals, temporal and spatial consistency of signals, and signal salience. Due to the valuable implication for personnel selection and training, many researches focus on influence of individual difference on vigilance performance and subjective experience (Finomore et al. 2009; Matthews et al. 2014; Reinerman-Jones et al. 2011; Shaw et al. 2010). Dittmar et al. (1993) found perceptual sensitivity for critical signals favored males in the monitoring the repetitive presentation of spatial task, and females felt the spatial task significantly more frustrating, mentally demanding, and effortful than males. Considering the five factor model of personality proposed by Costa and McCrae (1992), introversion and consciousness have been found positively related to vigilance performance (Koelega 1992; Rose et al. 2002).

Despite the abundant research on impacts of individual difference on vigilance, no study has examined the impact of cultural backgrounds on vigilance. Culture is often defined by nationality, race and organization, and we focused on national culture in this study. Over 40-years cross-cultural psychology research has shown that a wide range of psychological processes are influenced by culture, including attention (Masuda et al. 2008), perception, memory (Rule et al. 2011), and decision making (Chu et al. 2005). Because sustained attention on the stimuli over prolonged periods of time is an essential component of vigilance task, it is suggested cultural difference in the attentional process may indicate differences in vigilance.

Prior cross-cultural studies have found difference in cognition and attention between East Asians and Westerners. First, the different thinking style would bring cultural variation in attention pattern between East Asians and Westerners. East Asians tend to develop a holistic strategy of thinking style that emphasizes the entire field and relationships between the focal object and the context, whereas Westerners tend to develop a focused strategy of thinking style that emphasizes focusing on focal objects and ignoring contextual information (Duffy and Kitayama 2007; Duffy et al. 2009). It was found that East Asians paid more attention to the context and that Westerners paid more attention to

the focal object (Chua et al. 2005; Masuda and Nisbett 2006). Second, East Asians are more field-dependent and Westerners are more field-independence (Ji et al. 2000). Field dependence – independence refers to the degree to which perception of an object is influenced by the background or surroundings in which it appears. Ji et al. (2000) used rod and frame test to measure field independence between East Asians and Westerners. They found East Asians were more susceptible to the background, and Westerners were more confident about their performance.

The cross-culture difference in cognition and attention has been confirmed by a series of studies using different paradigms, including cognitive experimentation, eye movement patterns, and neuroscience evidence (Masuda and Nisbett 2001; Masuda et al. 2008; Masuda and Nisbett 2006). Goto et al. (2010) used event-related potential index N400 to test cultural difference in the visual processing of meaning, with a task of detecting incongruities between background and foreground objects. The result suggested that Asians were processing the relationship between foreground and background objects to a greater degree than Westerners, which was consistent with hypothesized greater holistic processing among East Asians. Researches on field dependence have found that field-dependent people had greater difficulty in maintaining attention on specific sectors of information (Avolio et al. 1981; Guisande et al. 2007; Jolly and Reardon 1985). Cahoon (1970) found that field dependence was negatively correlated with vigilance performance in terms of higher false alarm rate and lower perceptual sensitivity.

Therefore, cultural background influences the attention allocation. East Asians tend to allocate attention in a broad region (holistic attention strategy), whereas Westerners tend to focus attention on focal objects or events (focused attention strategy). It was also suggested that field-dependent people tend to have a broader, less efficient focused attention on the ongoing task, whereas field-independent people tend to have a narrow and efficient focused attention. Vigilance task is characterized by requiring observers to maintain focused attention on the focal targets or objects for a prolonged period of time. Considering cultural difference on attention pattern, we expect that culture has an influence on vigilance. Therefore, we propose the following hypothesis:

Hypothesis 1. Americans perform better than Chinese in single vigilance task.

Hypothesis 2. Chinese have higher levels of perceived workload than Americans in single vigilance task.

Hypothesis 3. Chinese have higher levels of perceived fatigue than Americans in single vigilance task.

The purpose of this research is to examine whether cultural differences (Chinese VS Americans) exist in performance and subjective experience associated with the vigilance task. The performance includes accuracy and response time. The subjective experience includes workload and fatigue. This study used Continuous Performance Test (CPT), a typical and simple vigilance task, in the experiment to test the cultural difference. It was shown to have high test-retest reliability (Borgaro et al. 2003).

2 Methodology

2.1 Participants

Twenty Chinese students (14 males, 6 females; age range = 20–28 years, mean age = 24 years) and twenty American students (14 males, 6 females, age range = 19–28 years, mean age = 24 years) at Tsinghua University participated in the experiment. The American participants had stayed in China for an average time of 5.25 months (SD = 3.52). Thirty percent (30%) of them were exchange students from the United States, and the rest were students who studied Chinese language at Tsinghua University. All participants had normal or corrected-to-normal vision.

2.2 Procedure

All observers monitored the 23-min repetitive presentation of symbols centered on the screen without interruption. The symbols were constructed in 100-point type on a white background. In order to assess culture differences on sustained attention, it is necessary to use culture-free stimuli in the experiment. We adapted the original CPT task by replacing the stimuli letters with a series of mathematical symbols (*e.g.* “ \perp ”, “ \wedge ”, “ \times ”, “ \pm ”, “ \angle ”, “ $=$ ”, “ $+$ ”). A pilot test was conducted to confirm a suitable frequency of stimuli in order to avoid the ceiling effect of performance. As a result, each stimulus exposed for 550 ms and there was no interval between stimuli in the experiment.

For each observer, the order of presentation of stimuli was varied at random with the target signal (angle symbol, “ \angle ”) occurred at a probability of $p = 0.25$. Observers signified their detection of target signals by pressing the spacebar key on a keyboard directly in front of them. Responses occurring within 550 ms during the presentation of a target signal were recorded as correct detections (hits). No response was required for other stimuli. All other responses were recorded as errors of commission (false alarms). Once the observer responded to a target signal or non-target signal, the stimulus disappeared immediately and the next stimulus appeared.

Perceived workload and fatigue were measured by a paper-pencil version of NASA-TLX and fatigue scale (Matthews and Desmond 1998). The scales were administered immediately following the CPT task. The NASA-TLX is a well-regarded instrument for the measurement of perceived mental workload (Hart and Staveland 1988). It provided a global measure of workload on a scale from 0 to 20 and also identified the relative contributions of six sources of workload: mental demand, temporal demand, physical demand, performance, effort, and frustration. Participants first provided ratings on the six subscales, and then conducted pair-wise comparisons to determine the relative importance of each subscale to the global workload score. The fatigue scale was an adaption of the scale developed by Matthews and Desmond (1998), which was designed to assess the fatigue induced by a vigilance task. After deleting some items unrelated to CPT task, fourteen items remained with three dimensions of fatigue, including perceptual fatigue, boredom and physical fatigue.

3 Results

3.1 Performance

Percentage of Correct Detections. Mean percentage of correct detections for Chinese and Americans are plotted as a function of periods of watch in Fig. 1. It is evident in the figure that the detection scores were generally quite high, exceeding 80% in all cases. The overall detection probability was greater for Americans ($M = 89.8\%$) than Chinese ($M = 82.4\%$), and moreover both groups demonstrated a vigilance decrement over time. The figure shows that there was a performance advantage favoring Americans and that detection probability in both groups declined over the duration of the vigilance task. These observations were supported by an analysis of variance (ANOVA), 2 (cultural background) \times 7 (period) mixed ANOVA, conducted on an arcsine transformation of the detection scores. In this analysis, and all subsequent CPT task analyses, the data was analyzed as seven sequential periods of work. Box's epsilon was used to correct the degrees of freedom if the Sphericity assumption was violated. As a result, it was found that the main effect for cultural background was significant, $F_{1,38} = 4.91$, $p = .033$, $\eta_p^2 = .114$, but the main effect of period and the interaction between these factors were not significant ($p > .05$).

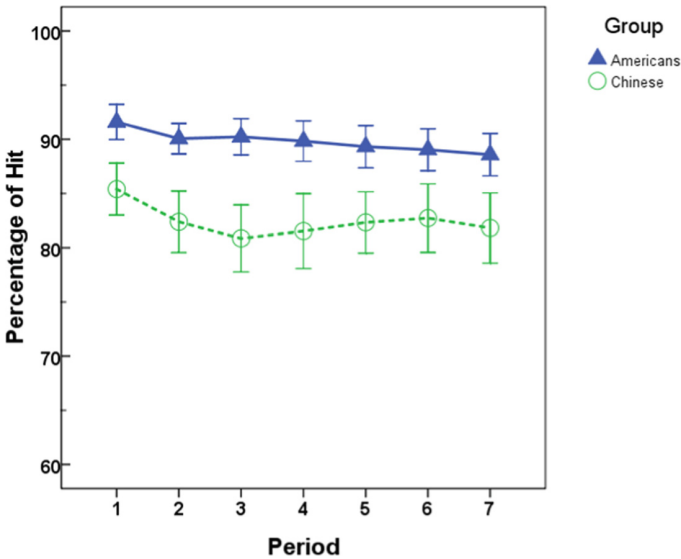


Fig. 1. Percentage of correct detections for Americans and Chinese

Percentage of False Alarm. False alarms were rare in this study. The mean percentage of false alarm was 1.7% for Americans and 2.0% for Chinese, respectively. A 2 (cultural background) \times 7 (period) mixed ANOVA analysis, based on the square root transformation of false alarm score, revealed a significant interaction effect between

these factors, as shown in Fig. 2, $F_{5.5,211} = 2.29$, $p = .041$, $\eta_p^2 = .057$. There was no significant main effect for cultural background or period. It can be seen in Fig. 2 that percentage of false alarm for Chinese are greater Americans in the previous periods while it fluctuations in the later periods.

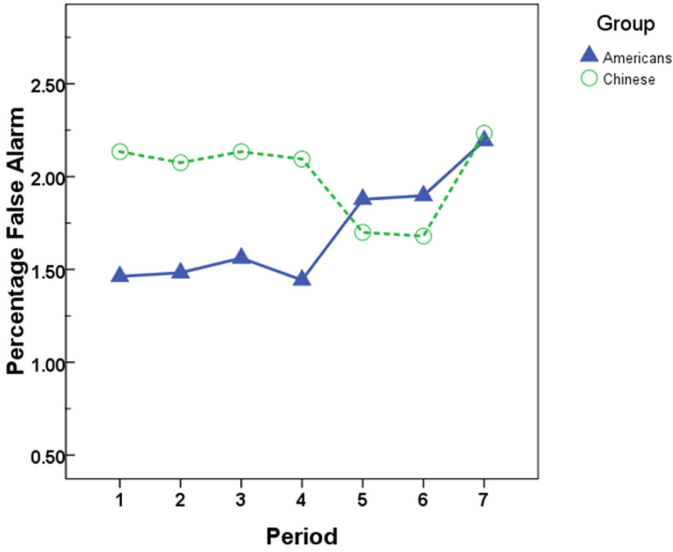


Fig. 2. Percentage of false alarm for Americans and Chinese

Reaction Time. Mean reaction time for Chinese and Americans are plotted as a function of periods of watch in Fig. 3. The overall reaction time was 400.36 ms for Americans and 414.72 ms for Chinese. The figure also shows that there was a performance advantage favoring Americans. These observations were supported by an analysis of variance (ANOVA), 2 (cultural background) \times 7 (period) mixed ANOVA. Significant main effects were observed for culture, $F_{1,38} = 7.21$, $p = .011$, $\eta_p^2 = .159$, and period, $F_{5.4,204} = 3.30$, $p = .006$, $\eta_p^2 = .080$, via a 2 (cultural background) \times 7 (period) mixed ANOVA analysis. However, the interaction effect between these factors was not significant, as shown in Fig. 3.

3.2 Subjective Experience

Workload. The mean (and standard errors) for global workload scores and six dimensions on the NASA-TLX for Chinese and Americans are displayed in Table 1. Cultural differences on these six dimensions were examined with t-test, or Mann-Whitney U test, if the normality assumption was violated. The global workload of Chinese was marginally higher than that of Americans, $t(38) = -1.88$, $p = .068$, $d = -0.59$. In terms of dimensions, Chinese rated significantly higher than Americans

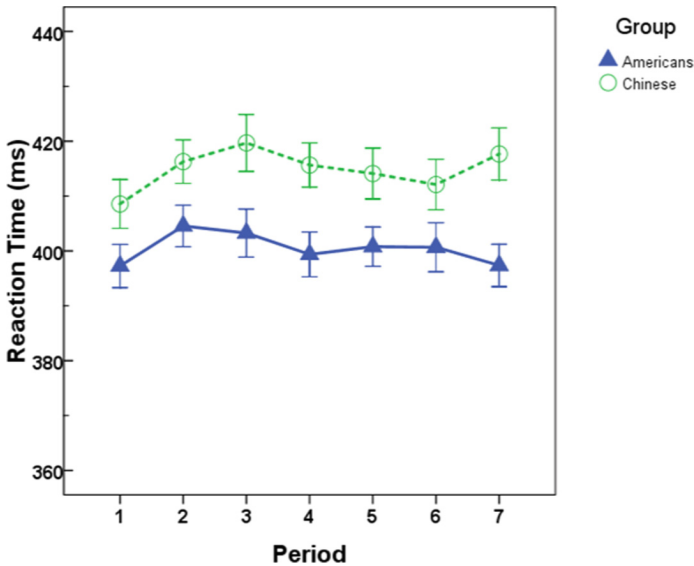


Fig. 3. Reaction time for Americans and Chinese

on physical demand ($U = 79.0$, $p = .001$, $r = 0.53$) and effort ($U = 122$, $p = .035$, $r = 0.33$). Chinese participants felt higher levels of physical demand and paid more efforts than American participants. For both Americans and Chinese, the overall workload on the two tasks was substantial.

Table 1. Ratings of NASA-TLX scales for Americans and Chinese

| Dimensions | Americans | Chinese | Statistics | p-value |
|-----------------|---------------|---------------|-----------------|---------|
| Mental demand | 5.98 (6.91) | 5.92 (7.58) | $U = 168.5$ | 0.40 |
| Physical demand | 2.97 (6.15) | 8.73 (7.37) | $U = 79.0$ | 0.001** |
| Temporal demand | 17.57 (7.42) | 14.47 (9.97) | $t(38) = 1.12$ | 0.27 |
| Performance | 7.35 (6.11) | 7.53 (7.23) | $U = 195.5$ | 0.90 |
| Effort | 8.38 (5.96) | 14.83 (9.13) | $U = 122$ | 0.04* |
| Frustration | 5.42 (8.75) | 4.83 (7.58) | $U = 199.5$ | 0.99 |
| Global workload | 47.67 (14.02) | 56.32 (15.09) | $t(38) = -1.88$ | 0.07* |

Note: Male-1, female-2. *Significant at 0.1 level; *Significant at 0.05 level; **Significant at 0.01 level.

Fatigue. The mean (and standard errors) for three dimensions of fatigue scales for Chinese and Americans are displayed in Table 2. The reliability was analyzed by Cronbach's alpha. For perceptual dimension, Cronbach's alpha of Chinese was 0.912 and Cronbach's alpha of Americans was 0.815. For physical dimension, Cronbach's alpha of Chinese was 0.813 and Cronbach's alpha of Americans was 0.618. For boredom dimension, Cronbach's alpha of Chinese was 0.916 and Cronbach's alpha of

Americans was 0.819. Cultural differences on these three dimensions were examined with t-test, or Mann-Whitney U test, if the normality assumption was violated. Other than perceptual fatigue, Chinese rated significantly higher than Americans on physical demand ($U = 122.5$, $p = .035$) and boredom ($U = 121.5$, $p = .033$). Chinese participants felt higher levels of physical fatigue and boredom than American participants.

Table 2. Ratings of fatigue scales for Americans and Chinese

| Dimensions | Americans | Chinese | Statistics | p-value |
|------------|-------------|-------------|-----------------|---------|
| Perceptual | 2.46 (1.01) | 2.28 (1.22) | $t(38) = 0.508$ | 0.614 |
| Physical | 0.71 (0.70) | 1.31 (0.97) | $U = 122.5$ | 0.035* |
| Boredom | 2.81 (0.99) | 2.00 (1.49) | $U = 121.5$ | 0.033* |

Note: Male-1, female-2. *Significant at 0.1 level; †Significant at 0.05 level; **Significant at 0.01 level.

4 Discussion

We developed a typical and basic continuous performance task (CPT) to examine the effect of cultural background (Chinese and Americans) on performance and subjective experience in vigilance task. Generally speaking, Americans performed better than Chinese in the experiment. The overall detection probability was greater for Americans ($M = 89.8\%$) than Chinese ($M = 82.4\%$). The percentage of correct detections generally showed a downward trend, which was consistent with vigilance decrement. However, only the main effect of cultural background was significant on percentage of correct detections ($F_{1,38} = 4.91$, $p = .033$, $\eta_p^2 = .114$). The main effect of period was not significant. A pilot test was conducted to confirm a suitable frequency of stimuli. Each stimulus exposed for 550 ms and there was no interval between stimuli in the experiment. The rate of stimuli in previous studies of CPT vigilance task was 60/min or less (Rosenberg et al. 2013; Temple et al. 2000). However, the task in this study was very simple because there is no mask covered the entire visual field and it was easy to differentiate target signals from non-target signals. A rate of 60/min result in ceiling effect since all target signals could be detected correctly. Therefore, stimulus exposed for 550 ms as a result of a pilot test. Previous studies of longer-duration CPTs have found a variety of changes in performance over time: While some have observed decrements over 30 min (Grier et al. 2003; Helton et al. 2005). The CPT vigilance task lasted about 23 min in this study. It showed a downward trend of correct detections and vigilance decrement may be significant during a longer period.

Besides accuracy, Americans responded to target signals faster than Chinese. The overall reaction time was 400.36 ms for Americans and 414.72 ms for Chinese. The main effect of cultural background was significant on reaction time ($F_{1,38} = 7.21$, $p = .011$, $\eta_p^2 = .159$). However, the main effect of period was not significant on reaction time. The reason may be the same as the percentage of correct detections. Because the task in this study was simple and duration of the task was relatively short, vigilance may not obviously decline. For percentage of false alarm, the main effects of

cultural background and period were not significant. Percentage of false alarm was rare in this study because it was 1.7% for Americans and 2.0% for Chinese. It has little effect on performance evaluation in vigilance task. Therefore, the result of CPT in this study showed that Chinese participants missed more target signals and responded more slowly than American participants.

This result can be explained by difference in attention allocation between East Asians and Westerners. East Asians tend to develop a holistic strategy of attention and Westerners tend to develop a focused strategy of attention (Duffy and Kitayama 2007; Duffy et al. 2009). Moreover, East Asians are more field-dependent and Westerners are more field-independence (Ji et al. 2000). East Asians tend to allocate attention in a broad region, whereas Americans are inclined to focus attention on focal objects or events (Chua et al. 2005; Masuda and Nisbett 2006). The vigilance task is characterized by requiring observers to maintain focused attention on the focal targets or objects for a prolonged period of time. The divided attention strategy of East Asians may lead them to be more vulnerable to vigilance failures than Americans, who are dominated by focused attention strategy.

In terms of six dimensions of NASA-TLX scales, Chinese rated significantly higher than Americans on physical demand ($U = 79.0$, $p = .001$, $r = 0.53$) and effort ($U = 122$, $p = .035$, $r = 0.33$). Chinese participants felt higher levels of physical demand and paid more efforts than American participants. For three dimensions of fatigue scales, Chinese rated significantly higher than Americans on physical demand ($U = 122.5$, $p = .035$) and boredom ($U = 121.5$, $p = .033$). Chinese participants felt higher levels of physical fatigue and boredom than American participants. In CPT vigilance task, Chinese participants felt more physical demand and higher levels of physical fatigue. Moreover, Chinese participants paid more effort to complete the CPT vigilance task but found they felt higher levels of boredom, loss of motivation during the experiment.

The subjective experience in the task was consistent with the explanation of difference in attention allocation between East Asians and Westerners. East Asians tend to allocate attention in a broad region and Americans are inclined to focus attention on focal objects or events. During the vigilance task, observers are required to focus on the focal object for a long period. According to mindlessness theory proposed by Robertson and his colleagues (Manly et al. 1999; Robertson et al. 1997), the repetitive and tedious nature of vigilance tasks leads the observers to disengage the attention from the ongoing activity and approach it in a thoughtless manner. Using a modification of the standard vigilance paradigm, the authors demonstrated that the vigilance performance is primarily determined by the duration of time over which attention must be maintained on the tasks. Chinese tend to develop a holistic strategy of attention and they are used to switch attention within a board region. As a result, Chinese participants consumed more cognitive resources and physical demand to control their attention on the local object during the task but had higher levels of perceived fatigue at the end of the task.

5 Conclusions

5.1 Theoretical Implications

This study conducted a typical and basic Continuous Performance Test (CPT) to examine the effect of cultural background on vigilance task. Result of this study reveals the difference in patterns of attention allocation between Chinese and Americans. The vigilance task is an attention demanding assignment, which requires the observers to maintain their focus attention on targets for prolonged periods of time. American participants performed better than Chinese and they had lower levels of perceived workload and fatigue than Chinese. The difference in performance and subjective experience indicates that American participants are inclined to focus attention on focal objects and that Chinese participants tend to allocate attention in a broad region.

5.2 Practical Implications

Exploring the effects of culture on vigilance task is of great importance for practical applications of vigilance research. As noted by Chapanis (1974), “failure to take account of national and cultural variables may nullify the gains that one might reasonably expect to follow from the application of ergonomics in many areas of the world”. The personnel selection and assessment criteria for vigilant operators, which were primarily derived from Western cultures, may not be appropriate for Eastern cultures. Moreover, the automated systems which are designed according to Western minds may do not fit Eastern people. In the field of aviation, especially for cockpit operations, accumulating data indicated that there are substantial cultural differences in the way pilots conducts their work. Then, the researchers suggested that the training for crew resource management, which includes situation awareness and vigilance, should be adapted to national culture so as to make it more effective (Helmreich et al. 2001).

5.3 Limitations and Future Research

A major limitation of this study is controlling confounding variables that may influence performance and subjective experience in the task. Impact of individual differences on vigilance are not controlled in the experiment. In addition, the duration of the task is a little short since the main effect of period was not significant on correct detections and reaction time. Vigilance decrement is a major finding of prior vigilance research and longer duration may showed an obvious decrement of performance because the task in the experiment is very simple and basic.

The future research may be examine the effect of cultural background on performance and subjective experience in multiple vigilance tasks. Based on finding that East Asians tend to allocate attention in a broad region and Americans are inclined to focus attention on focal objects or events in this study, we expect that Chinese perform better than Americans in multiple vigilance task.

Acknowledgments. This study was supported by the National Natural Science Foundation of China (Project no. 71671102).

References

- Avolio, B.J., Alexander, R.A., Barrett, G.V., Sterns, H.L.: Designing a measure of visual selective attention to assess individual differences in information processing. *Appl. Psychol. Meas.* **5**(1), 29–42 (1981)
- Baker, K., Olson, J., Morisseau, D.: Work practices, fatigue, and nuclear power plant safety performance. *Hum. Factors* **36**(2), 244–257 (1994)
- Borgaro, S., Pogge, D.L., DeLuca, V.A., Bilginer, L., Stokes, J., Harvey, P.D.: Convergence of different versions of the continuous performance test: clinical and scientific implications. *J. Clin. Exp. Neuropsychol.* **25**(2), 283–292 (2003)
- Cahoon, R.L.: Vigilance performance under hypoxia. *J. Appl. Psychol.* **54**(6), 479–483 (1970)
- Chapanis, A.: National and cultural variables in ergonomics†. *Ergonomics* **17**(2), 153–175 (1974)
- Chu, P.C., Spires, E.E., Farn, C.K., Sueyoshi, T.: Decision processes and use of decision aids: comparing two closely related nations in East Asia. *J. Cross Cult. Psychol.* **36**(3), 304–320 (2005)
- Chua, H.F., Boland, J.E., Nisbett, R.E.: Cultural variation in eye movements during scene perception. *Proc. Natl. Acad. Sci. U.S.A.* **102**(35), 12629–12633 (2005)
- Costa, P.T., McCrae, R.R.: The five-factor model of personality and its relevance to personality disorders. *J. Pers. Disord.* **6**(4), 343–359 (1992)
- Dillard, M.B., Warm, J.S., Funke, G.J., Funke, M.E., Finomore, V.S., Matthews, G., Shaw, T.H., Parasuraman, R.: The sustained attention to response task (SART) does not promote mindlessness during vigilance performance. *Hum. Factors* **56**(8), 1364–1379 (2014)
- Dittmar, M.L., Warm, J.S., Dember, W.N., Ricks, D.F.: Sex differences in vigilance performance and perceived workload. *J. Gen. Psychol.* **120**(3), 309–322 (1993)
- Drews, F.A.: Human factors in critical care medical environments. *Rev. Hum. Factors Ergon.* **8**(1), 103–148 (2013)
- Duffy, S., Kitayama, S.: Mnemonic context effect in two cultures: attention to memory representations? *Cogn. Sci.* **31**(6), 1009–1020 (2007)
- Duffy, S., Toriyama, R., Itakura, S., Kitayama, S.: Development of cultural strategies of attention in North American and Japanese children. *J. Exp. Child Psychol.* **102**(3), 351–359 (2009)
- Finomore, V., Matthews, G., Shaw, T., Warm, J.: Predicting vigilance: a fresh look at an old problem. *Ergonomics* **52**(7), 791–808 (2009)
- Funke, G.J., Warm, J.S., Baldwin, C.L., Garcia, A., Funke, M.E., Dillard, M.B., Finomore Jr., V. S., Matthews, G., Greenlee, E.T.: The independence and interdependence of coaching observers in regard to performance efficiency, workload, and stress in a vigilance task. *Hum. Factors* **58**(6), 915–926 (2016)
- Gillberg, M., Åkerstedt, T.: Sleep loss and performance: no “safe” duration of a monotonous task. *Physiol. Behav.* **64**(5), 599–604 (1998)
- Goto, S.G., Ando, Y., Huang, C., Yee, A., Lewis, R.S.: Cultural differences in the visual processing of meaning: detecting incongruities between background and foreground objects using the N400. *Soc. Cogn. Affect. Neurosci.* **5**(2–3), 242–253 (2010)
- Grier, R.A., Warm, J.S., Dember, W.N., Matthews, G., Galinsky, T.L., Szalma, J.L., Parasuraman, R.: The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Hum. Factors* **45**(3), 349–359 (2003)
- Guisande, M.A., Páramo, M.F., Tinajero, C., Almeida, L.S.: Field dependence-independence (FDI) cognitive style: an analysis of attentional functioning. *Psicothema* **19**(4), 572–577 (2007)

- Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) *Advances in Psychology*, vol. 52, pp. 139–183. North-Holland, Amsterdam (1988)
- Hawley, J., Mares, A., Giammanco, C.: Training for effective human supervisory control of air and missile defense systems. US Army Research Laboratory (2006)
- Helton, W.S., Hollander, T.D., Warm, J.S., Matthews, G., Dember, W.N., Wallaart, M., Beauchamp, G., Parasuraman, R., Hancock, P.A.: Signal regularity and the mindlessness model of vigilance. *Br. J. Psychol.* **96**(2), 249–261 (2005)
- Helmreich, R.L., Wilhelm, J.A., Klinec, J.R., Merritt, A.C.: Culture, error, and crew resource management. In: Salas, E., Bowers, C.A., Edens, E. (eds.) *Improving Teamwork in Organizations: Applications of Resource Management Training*, pp. 305–331. Lawrence Erlbaum Associates, Mahwah (2001)
- Helton, W.S., Russell, P.N.: The effects of arousing negative and neutral picture stimuli on target detection in a vigilance task. *Hum. Factors* **53**(2), 132–141 (2011)
- Ji, L., Peng, K., Nisbett, R.E.: Culture, control, and perception of relationships in the environment. *J. Pers. Soc. Psychol.* **78**(5), 943–955 (2000)
- Jolly, E.J., Reardon, R.: Cognitive differentiation, automaticity, and interruptions of automatized behaviors. *Pers. Soc. Psychol. Bull.* **11**(3), 301–314 (1985)
- Kamzanova, A.T., Kustubayeva, A.M., Matthews, G.: Use of EEG workload indices for diagnostic monitoring of vigilance decrement. *Hum. Factors* **56**(6), 1136–1149 (2014)
- Koelega, H.S.: Extraversion and vigilance performance: 30 years of inconsistencies. *Psychol. Bull.* **112**(2), 239–258 (1992)
- Langan-Fox, J., Sankey, M.J., Cauty, J.M.: Human factors measurement for future air traffic control systems. *Hum. Factors* **51**(5), 595–637 (2009)
- Manly, T., Robertson, I.H., Galloway, M., Hawkins, K.: The absent mind: further investigations of sustained attention to response. *Neuropsychologia* **37**(6), 661–670 (1999)
- Masuda, T., Nisbett, R.E.: Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans. *J. Pers. Soc. Psychol.* **81**(5), 922–934 (2001)
- Masuda, T., Gonzalez, R., Kwan, L., Nisbett, R.E.: Culture and aesthetic preference: comparing the attention to context of East Asians and Americans. *Pers. Soc. Psychol. Bull.* **34**(9), 1260–1275 (2008)
- Masuda, T., Nisbett, R.E.: Culture and change blindness. *Cogn. Sci.* **30**(2), 381–399 (2006)
- Matthews, G., Desmond, P.A.: Personality and multiple dimensions of task-induced fatigue: a study of simulated driving. *Pers. Individ. Differ.* **25**(3), 443–458 (1998)
- Matthews, G., Warm, J.S., Shaw, T.H., Finomore, V.S.: Predicting battlefield vigilance: a multivariate approach to assessment of attentional resources. *Ergonomics* **57**(6), 856–875 (2014)
- Molloy, R., Parasuraman, R.: Monitoring an automated system for a single failure: vigilance and task complexity effects. *Hum. Factors* **38**(2), 311–322 (1996)
- Parasuraman, R.: Vigilance, monitoring, and search. In: Boff, K.R., Kaufman, L., Thomas, J. P. (eds.) *Handbook of Perception and Human Performance. Cognitive Processes and Performance*, vol. 2, pp. 1–39. Wiley, Oxford (1986)
- Parasuraman, R., Davies, D.R.: A taxonomic analysis of vigilance performance. In: Mackie, R.R. (ed.) *Vigilance*, pp. 559–574. Springer, Boston (1977). https://doi.org/10.1007/978-1-4684-2529-1_26
- Reinerman-Jones, L.E., Matthews, G., Langheim, L.K., Warm, J.S.: Selection for vigilance assignments: a review and proposed new direction. *Theor. Issues Ergon. Sci.* **12**(4), 273–296 (2011)

- Robertson, I.H., Manly, T., Andrade, J., Baddeley, B.T., Yiend, J.: 'Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia* **35**(6), 747–758 (1997)
- Rose, C.L., Murphy, L.B., Byard, L., Nikzad, K.: The role of the big five personality factors in vigilance performance and workload. *Eur. J. Pers.* **16**(3), 185–200 (2002)
- Rosenberg, M., Noonan, S., DeGutis, J., Esterman, M.: Sustaining visual attention in the face of distraction: a novel gradual-onset continuous performance task. *Atten. Percept. Psychophys.* **75**(3), 426–439 (2013)
- Rule, N.O., Freeman, J.B., Ambady, N.: Brain, behavior, and culture: insights from cognition, perception, and emotion. In: Han, S., Pöppel, E. (eds.) *Culture and Neural Frames of Cognition and Communication*, pp. 109–122. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-15423-2_7
- Shaw, T.H., Matthews, G., Warm, J.S., Finomore, V.S., Silverman, L., Costa, P.T.: Individual differences in vigilance: personality, ability and states of stress. *J. Res. Pers.* **44**(3), 297–308 (2010)
- Sheridan, T.: *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, Cambridge (1992)
- Sheridan, T.B.: *Humans and Automation: System Design and Research Issues*. Wiley, New York (2002)
- Smallwood, J., Davies, J.B., Heim, D., Finnigan, F., Sudberry, M., O'Connor, R., Obonsawin, M.: Subjective experience and the attentional lapse: task engagement and disengagement during sustained attention. *Conscious. Cogn.* **13**(4), 657–690 (2004)
- Temple, J.G., Warm, J.S., Dember, W.N., Jones, K.S., LaGrange, C.M., Matthews, G.: The effects of signal salience and caffeine on performance, workload, and stress in an abbreviated vigilance task. *Hum. Factors* **42**(2), 183–194 (2000)
- Warm, J.S., Parasuraman, R., Matthews, G.: Vigilance requires hard mental work and is stressful. *Hum. Factors* **50**(3), 433–441 (2008)
- Young, M.S., Stanton, N.A.: What's skill got to do with it? Vehicle automation and driver mental workload. *Ergonomics* **50**(8), 1324–1339 (2007)



The Impact of Metacognitive Monitoring Feedback on Mental Workload and Situational Awareness

Jung Hyup Kim^(✉)

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

Abstract. The need to develop more effective feedback has become a growing concern in training. Feedback should be designed to provide meaningful information in order to help them improve their performance. On the other hand, the feedback should be designed not to increase the learners' mental workload even while they maximize the benefits of using such feedback during training. Recently, Kim [1] developed the metacognitive monitoring feedback method. This methodology was tested in a computer-based training environment. The authors' results showed that metacognitive monitoring feedback significantly improved participants' performance during two days of a training session. However, the previous study did not investigate the impact of metacognitive monitoring feedback on participants' mental workload and situational awareness. Hence, in this study, we investigated those needs and found a negative relationship between situational awareness and workload when the trainees observed the metacognitive monitoring feedback.

Keywords: Metacognition · Mental workload · Situational awareness

1 Introduction

Developing advanced training methods that not only consider situational awareness but also mental workload is a growing concern in a computer-based training environment. According to Norman [2], learners' situational awareness can be improved when trainees observe feedback that contains valuable information related to the task they learn during a training session. Many studies have been conducted to develop a better training method that can improve trainees' situational awareness without increasing their workload. Among them, the concept of metacognitive monitoring feedback was recently developed by Kim [1]. The feedback showed a significant performance improvement on a visual identification task in a computer-based training environment [1]. It was post-test feedback and showed the central role of trainees' learning in a human-in-the-loop simulation. However, how the metacognitive monitoring feedback affects trainees' mental workload was not tested in the previous studies. Hence, the effect of metacognitive monitoring feedback on mental workload was investigated in the current study. Also, it is important to understand how metacognitive monitoring feedback influences the relationship

between mental workload and situational awareness. For this reason, we studied this relationship resulting from the metacognitive monitoring feedback. In this study, Endsley's situation awareness model [3] was used as the underlying basis for measuring trainees' situational awareness. Also, retrospective confidence judgments (RCJ) and the NASA task load index (TLX) were used as metrics for assessing confidence levels and mental workload, respectively.

For the experimental group, SA level-based metacognitive monitoring feedback was designed for the experiment. The participants were exposed to the feedback screens after they had answered all situation awareness probes. The participants were monitored for a percentage of their responses for each level of SA, and each level of retrospective judgment was rated separately. The participants who were assigned to the control group did not receive the metacognitive monitoring feedback after they answered all SA questions.

The primary research questions were as follows: Does SA level-based metacognitive monitoring feedback influence learner workload? If so, is there any correlation between situational awareness and workload? The following hypotheses were tested in the human-in-the-loop simulation environment.

- Hypothesis #1: The metacognitive monitoring feedback significantly influences learner workload.
- Hypothesis #2: The metacognitive monitoring feedback significantly influences the correlation between situational awareness and workload.

2 Method

2.1 Computer-Based Training Environment

To test both hypotheses, a time windows-based human-in-the-loop simulation was used as a training tool. In this simulation framework, every event generated from the simulator was based on the concept of time window developed by Thiruvengada and Rothrock [4]. During the training, participants were required to learn how to defend their battleship against hostile aircraft. Their main task was identifying unknown aircraft and taking appropriate actions. To defend the ship, they must learn the Rules of Engagement (RoE). To recognize the identification of unknown aircraft, they need to understand the meaning of cues that related to the identification of the aircraft. Figure 1 shows the simulation interface, and the details of the Rules of Engagement are shown in Table 1.

Table 1. Rules of Engagement for the radar monitoring task

| Rules of Engagement | Descriptions |
|---------------------|--|
| Identification | Make a primary identification and AIR identification - Primary identification: friendly or hostile - AIR identification: Strike, Missile, Helicopter, Commercial Air, Airborne Early Warning |

(continued)

Table 1. (continued)

| | |
|---------------------|---|
| Rules of Engagement | Descriptions |
| Warning | Issue three levels of warning - Level 1: Issue first warning (50–40 NM) - Level 2: Issue second warning (40–30 NM) - Level 3: Issue final warning at 30 NM |
| Assign | Engage the target aircraft (less than 30 NM) |

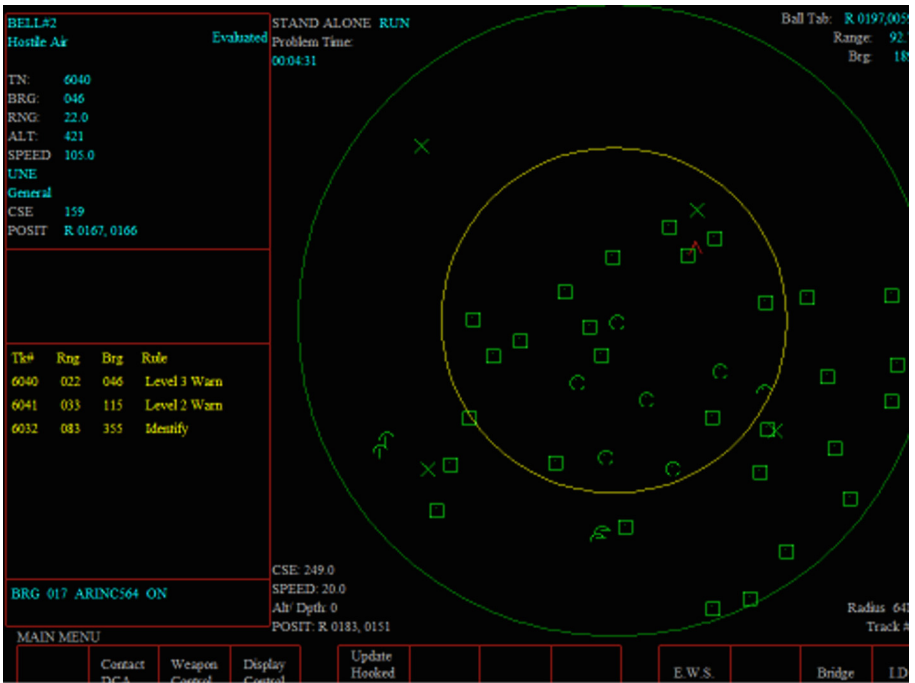


Fig. 1. Time windows-based human-in-the-loop simulation interface

2.2 Procedure

The experiment consisted of two sessions – a practice session and training session. Before the experiment, the participants were asked their previous experience with a radar monitoring task and video game. The participants took a 60 min practice session. During this session, the participants learned task-specific skills, such as how to perform the rules of engagement, how to identify unknown aircraft, and how to engage the target aircraft. After that, they received an instructor’s feedback about their performance. They also experience several practice runs during the session. Each practice scenario took 5 min to complete. Figure 2 shows the detail procedure for the practice session.

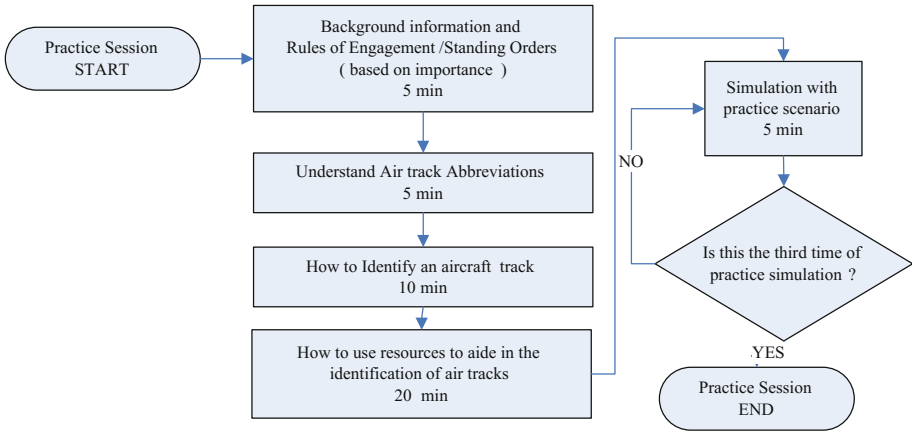


Fig. 2. The practice session procedure

The participants underwent a training session. Figure 3 shows the detailed procedure of the training session.

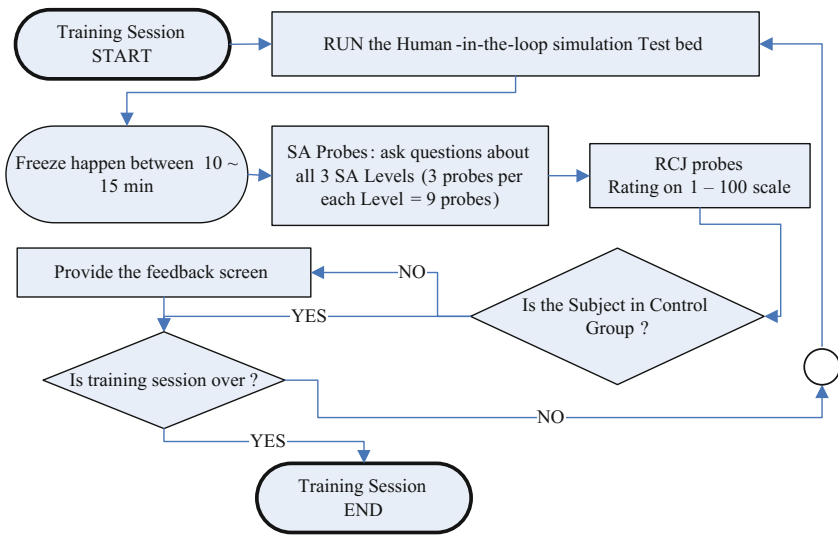


Fig. 3. The training session procedure

Each participant performed multiple different scenarios. Table 2 shows one of the simulation scenarios that was used in this study. In this scenario, the total number of aircraft and time windows were twenty-four and forty-six, respectively. There were four friendly aircraft and six known aircraft (start TN with 40**) and fourteen unknown aircraft (start TN with 60**) in the scenario. All events occurred based on time windows in specific time sequences and were tied to situational awareness questionnaires.

Table 2. Description of simulation scenario [5]

| Time | Event | Event Description | Detail |
|--------|--|---|---|
| 0:06 | initialization | appear unidentified aircrafts | -Known TN4009, 4010, 4011, 4012:DCA TN4003, 4004, 4005 : friendlyStrike TN4006, 4007: Commercial aircraft |
| | observation | give time to participants to observe all flying object to make them understand the situation | -Unknown TN6015, 6016, 6017, 6018: Commercial TN6019, 6020, 6021: Hostile Strike TN6022: Hostile Strike |
| 2:18 | Spying activity | Behavior of unknown aircraft which has spying purpose: aircraft from <i>Chodian</i> attempt to cross the border into <i>Koraban</i> | -Unknown TN6023: AEW appear and Move to EAST |
| 4:08 | Abnormal activity | abnormal behavior: malfunction of private aircrafts from <i>Koraban</i> | -Unknown TN6024 appear and move to <i>Irascibal</i> border |
| 4:35 | Unidentified International Commercial Aircraft | international commercial aircraft without IFF (Identification Friend or Foe) information observe in “Flight Air Route” | -Unknown TN6026 appear and move to Jovania international Airport in <i>Chodian</i> |
| 5:02 | Unidentified International Commercial Aircraft | international commercial aircraft without IFF (Identification Friend or Foe) information observe in “Flight Air Route” | -Unknown TN6027 move to Genialistan international Airport in <i>Korban</i> |
| 5:51 | Spying Activity | Behavior of unknown aircraft which has spying purpose: Hostile aircraft attempt to cross the border into <i>Koraban</i> | -Unknown TN6019, 6020 (Hostile Strike) are slow down their speed and altitude |
| 7:12 | Hostile Activity | Pop up the additional unidentified aircraft within 50NM | -Unknown TN6028 (Hostile Strike) appear on radar Screen |
| 7:53 | Practice fire from Hostile | fire missile to practice target in the hostile territory (Training purpose) | -Unknown TN6022, 6028 (Hostile Strike) fire Missile |
| 10:11 | Practice fire from Friendly | fire missile to practice target in the friendly territory (Training purpose) | -Known TN4003, 4004 (friendly strike) fire Missile |
| 13: 29 | Execute SA Probe | Pause simulation and execute SA Probe | |

While the participants were performing one of the training scenarios, the simulation was frozen automatically at a random time between 10 and 15 min. After the freeze, the participants saw the screens for SA probes and RCJ probes.

The following are examples of situation awareness questionnaires used in the experiment:

- Level 1 SA: Perception of the Elements in the Environment
 - **Question:** What was the location of TN6026?
 - **Choice:** Within 50 NM, 40 NM, 30 NM, or 20 NM
- Level 2 SA: Comprehension of the Current Situation
 - **Question:** What was the primary identification of TN6023?
 - **Choice:** Friendly or Hostile
- Level 3 SA: Projection of Future Status
 - **Question:** TN6027 is following “Flight Air Route” and moving to “Genialistan”?
 - **Choice:** True or False

The following are the probes for RCJ:

- RCJ probes based on SA Level
 - **Level 1:** “How well do you think you have detected the objects in your airspace?”
 - **Level 2:** “How well do you think you are aware of the current overall situation of your airspace?”
 - **Level 3:** “How well do you think you are aware of where the overall situation of your airspace is heading?”

After they answered all questions, the participants in the experimental group received the SA level-based metacognitive monitoring feedback. The others in a control group did not receive any feedback. Figure 4 shows an example of the SA level-based metacognitive monitoring feedback. The feedback consists of three main components: (1) a screenshot of the frozen moment with the answers of SA questions; (2) Participant’s SA responses, correct SA answers, and SA questions; (3) Visual graphs of both RCJ and SA performance.

The participants in the experimental group observed the feedback screens that contain the information regarding how they answered all SA probes with the images of the radar monitor at the frozen moment and the results of each level of SA probes as well as each level of RCJ scores. The exposure time for the feedback screen was 1 min to minimize the effect of bias due to uneven exposure. The control group did not receive any feedback.

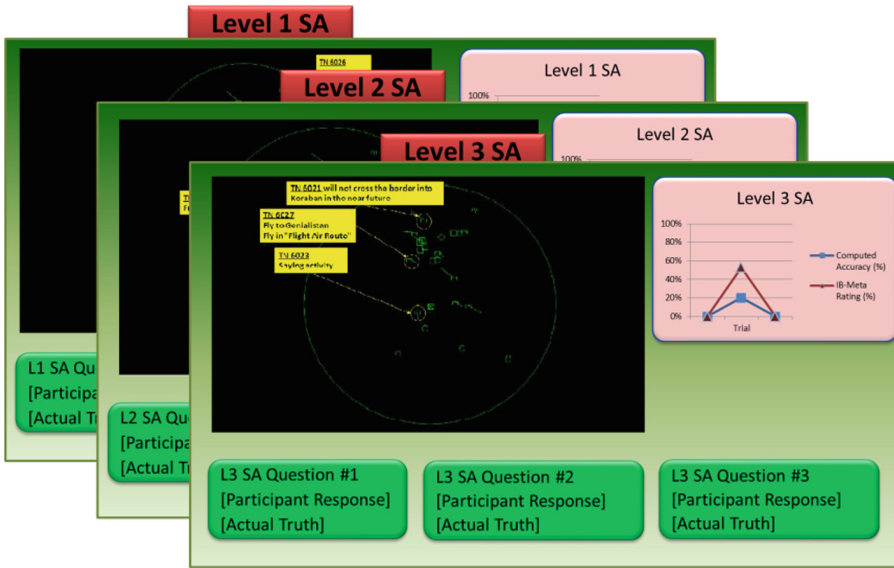


Fig. 4. SA level-based Metacognitive monitoring feedback

2.3 Performance Metrics

Retrospective Confidence Judgments (RCJ). Retrospective Confidence Judgments (RCJ) is one of the metacognitive monitoring metrics that is commonly used in research related to metacognition [6]. This is a self-rating report regarding the participants' confidence level for their responses before knowing whether they are correct or incorrect. RCJ is one of the ways to understand the metacognitive monitoring processes associated with retrieval of metamemory [7]. We collected RCJ scores (scale: 1 to 100) after the participants answered the SA probes during the testing sessions.

Situational Awareness Global Assessment Technique (SAGAT). SAGAT is one of the famous measures of situational awareness [8]. It is designed for a computer-based simulation environment in dynamic systems (e.g., driving simulator, flight simulator, or process monitoring task). This technique was used in this study to collect participants' situational awareness in given conditions. To measure their situational awareness, SA probes for each SA level were presented to the participants after a simulation clock passed 10 min from the beginning. The accuracy of participant's situational awareness (SA Accuracy) was calculated by

$$\text{SA accuracy} = \text{Total number of correct response} / \text{Total number of SA probes} \quad (1)$$

NASA-Task Load Index (TLX). It is the most well-known measure of subjective workload technique. NASA-TLX consists of mental demand, physical demand, temporal demand, performance, effort, and frustration. This multidimensional subjective

workload rating technique is commonly used as a tool to assess operator's workload related to aviation tasks [9] and flight simulators [10]. If NASA-TLX score is close to 100, it represents a high workload. If the score is close to 0, it means the operator had a low workload while he or she performed the task.

3 Results

3.1 Analysis of Variance

We compared participants' RCJ score, SA accuracy, and NASA-TLX score between the groups. For the RCJ, there were no significant differences between the groups; RCJ ($F(2,90) = 1.05$, $p = 0.357$). In addition, there was no significant difference on NASA-TLX between the groups ($F(2,90) = 0.16$, $p = 0.849$). However, SA accuracy was significantly different between the groups ($F(2,90) = 7.95$, $p < 0.001$). The experimental group's SA accuracy was significantly higher than the control group.

3.2 Correlation Matrix

Table 3 shows correlations between RCJ, SA, and NASA-TLX for both groups. The experimental group shows significant correlations between RCJ, SA accuracy, and NASA-TLX, while the control group shows a correlation between SA accuracy and RCJ (no correlation between RCJ and NASA-TLX and between SA accuracy and NASA-TLX).

Table 3. Correlation comparisons between the control group and experimental group.

| Measure | RCJ | | SA accuracy | |
|-------------|-----------------|----------------|-----------------|--------|
| | E | C | E | C |
| SA accuracy | 0.376** | 0.236** | - | - |
| NASA-TLX | -0.353** | -0.09 | -0.298** | -0.071 |

4 Discussion

The present study compared the effects of SA level-based metacognitive monitoring feedback on situational awareness in a computer-based training environment. The accuracy of situational awareness, mental workload, and subject ratings of retrospective confidence judgments were collected through the human-in-the-loop simulation.

- **Hypothesis #1:** The metacognitive monitoring feedback significantly influences learner workload.

NO, there was no evidence to support the hypothesis that SA level-based metacognitive feedback significantly affects trainees' mental workload when we compared the NASA-TLX scores between the two groups. Therefore, we infer that this metacognitive monitoring feedback does not increase the learners' workload during the

training. To understand this phenomenon, further analysis of NASA-TLX data between groups is necessary.

- **Hypothesis #2:** The SA-based metacognitive monitoring feedback significantly influences the correlation between situational awareness and workload.

YES, we found a negative correlation between SA accuracy and NASA-TLX in the experimental group. It shows that the participants who had better situational awareness experienced a lower mental workload during the training, while the performers with a poor situational awareness showed a higher mental workload. This phenomenon might be explained by the metacognitive framework developed by Nelson and Narens [11]. According to the framework, there are two layers in human cognition: (1) meta-level and (2) object-level. Here, object-level is defined as the process of cognitive activities from human sensors (e.g., vision, hearing, taste, smell, or touch). Meta-level is defined as a mental model of a particular task related to meta-knowledge from object-level. Many studies in the field of metacognition have shown that students could learn new concepts and skills through the interplay of these two levels, and the communication between these two levels plays one of the critical factors to stimulate student's learning process. During the experiment, the SA-based metacognitive monitoring feedback provided the latest meta-knowledge of the radar monitoring task and helped the trainees update their mental models of unknown aircraft identification. In other words, the participants could easily observe the modification of the meta-knowledge in object-level through the feedback. However, the control group was not able to receive the metacognitive monitoring feedback. Hence, they could not efficiently update their mental models of the identification task compared to the experimental group.

In this study, we investigated the impact of metacognitive monitoring feedback on mental workload and situational awareness in a computer-based training environment. The initial findings of our study provided a better understanding of the metacognitive monitoring process and its relation to workload in a computer-based environment.

There are several limitations of the present study. First, the experiment has not been formatted to interpret the underlying workload mechanism between object-level and meta-level. Hence, the future research should investigate the cognitive-affective status of the learners with their workload levels by using biosensors (e.g., electroencephalography, eye tracking, and electrocardiography). Secondly, the findings of this study are limited to visual identification tasks. For that reason, it would be better to investigate the effect of metacognitive monitoring feedback in different domains.

References

1. Kim, J.H.: The effect of metacognitive monitoring feedback on performance in a computer-based training simulation. *Appl. Ergon.* **67**, 193–202 (2018)
2. Norman, D.A.: The 'problem' with automation: inappropriate feedback and interaction, not 'over-automation'. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **327**(1241), 585–593 (1990)
3. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **37**(1), 32–64 (1995)

4. Thiruvengada, H., Rothrock, L.: Time windows-based team performance measures: a framework to measure team performance in dynamic environments. *Cogn. Technol. Work* **9** (2), 99–108 (2007)
5. Kim, J.H.: *Developing a Metacognitive Training Framework in Complex Dynamic Systems Using a Self-regulated Fuzzy Index* (2013)
6. Dunlosky, J., Metcalfe, J.: *Metacognition*. Sage Publications, Thousand Oaks (2008)
7. Dougherty, M.R., et al.: Using the past to predict the future. *Mem. Cogn.* **33**(6), 1096–1115 (2005)
8. Endsley, M.R.: *Situation Awareness Global Assessment Technique (SAGAT)*. IEEE (1988)
9. Nygren, T.E.: Psychometric properties of subjective workload measurement techniques: implications for their use in the assessment of perceived mental workload. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **33**(1), 17–33 (1991)
10. Hancock, P., Williams, G., Manning, C.: Influence of task demand characteristics on workload and performance. *Int. J. Aviat. Psychol.* **5**(1), 63–86 (1995)
11. Nelson, T.O., Narens, L.: Metamemory: a theoretical framework and new findings. *Psychol. Learn. Motiv.* **26**, 125–141 (1990)



A Heterarchical Urgency-Based Design Pattern for Human Automation Interaction

Axel Schulte¹, Diana Donath¹, Douglas S. Lange^{2(✉)},
and Robert S. Gutzwiller²

¹ Universität der Bundeswehr München (UBM), Neubiberg, Germany
{axel.schulte,diana.donath}@unibw.de

² Space and Naval Warfare Systems Center Pacific (SPAWAR),
San Diego, CA, USA
{dlange,gutzwill}@spawar.navy.mil

Abstract. We document a Human-Autonomy Teaming design pattern to provide a means for an task management assistant to mitigate errors that may occur due to changes in urgency levels of tasks. Urgency can increase or decrease due to changes in the task environment, or through failure to begin execution of a task at the correct time. We discuss the structure and key aspects of the pattern and provide a sample implementation. We also discuss the key aspects of the human partner's performance that must be measured and considered in implementing such a pattern. Finally, we discuss known issues and other related patterns.

Keywords: Human-autonomy teaming · Pattern · Task management

1 Intent

Our intent is to provide a means for an assistant/associate system to mitigate erroneous behavior of an operator by a stepwise, increasing intervention/support. The interventions of the assistant/associate system range from alerts, messages, and suggestions, up to overrides in order to transition a dangerous situation into a normative (safe) one. The stepwise approach strives to keep the operator vigilant with respect to the task, and responsible for as long as possible for task accomplishment. Another objective of stepwise intervention is to avoid a degradation of the final work result, which may only be possible as long as any error caused by the human has no direct/immediate negative effect on the overall work objective. For this reason, we suggest a stepwise error correction only for errors which are still repairable before degradation is realized.

We consider two kinds of erroneous behaviors as related to human performance when interacting with automation:

1. Errors which occur when the human fails to take a necessary action (errors of omission) and;
2. Errors caused by a wrongly selected, wrongly executed, or improperly timed executed action (errors of commission).

The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.

These two types of errors are found in studies of human automation interaction [1] and are especially critical aspects of human interaction when automation may make decisions [2]. Intuitively, with increased time pressure, humans may be more likely to accept automation recommendations or rely on the system in order to conserve mental resources. Reliance on aids may increase as operators reach their task saturation limits [3], but it is more complicated in determining if urgency itself dictates this relationship. One study that manipulated time pressure showed no relationship with automated aid reliance in an air traffic control context [4]. Nevertheless, these results may not apply as they dealt in decisions to use the automation, whereas the current design pattern instantiates an automated solution automatically to keep damage or errors from occurring.

In many contexts, task urgency is highly related to safety outcomes. In driving, a key factor in accidents is time following some emergent information, such as a truck pulling out onto the highway – is there enough time to avoid it (a physical limitation), and what actions must be taken through to task completion within that amount of time to avoid a serious accident. Methods which “create” more time (through reduction in speed, heads-up alerts, etc.) are then generally successful at enabling the human to respond more effectively. In the figure below, this would represent pushing the threshold for a task completion (e.g., maneuvering away from the truck) further toward the right where the human has time to respond and the response will be effective (top Fig. 1). The system could also respond in this case, but as discussed above, may need to be left idle to ensure the human remains aware and engaged in driving. There remain situations on the leeward edge of the urgency “continuum,” in which so little time is available for a safety-critical task to be accomplished that full automation is used and justified (bottom Fig. 1). In that case, the system may even lock the human from responding altogether to avoid any interaction issues.

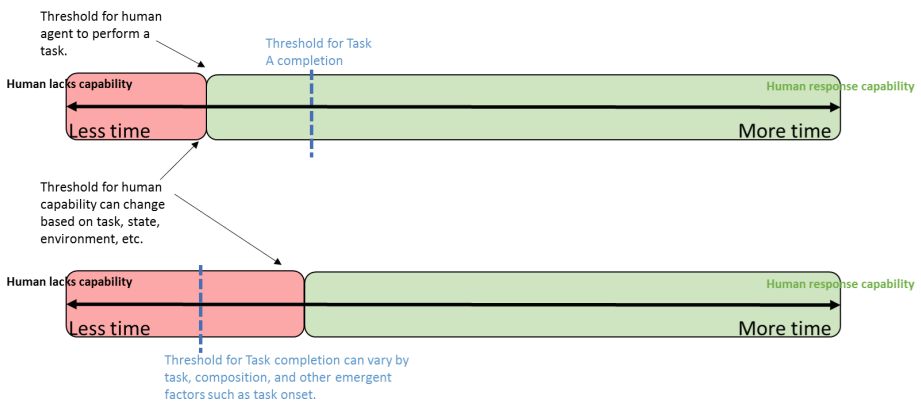


Fig. 1. Urgency continuum abstract representation to illustrate time limits on human capacity to perform tasks.

The current design pattern is better understood through examining the general cases – those in which intervention before error is possible by the human or the system. Other

domains that may make use of this continuum are in automated vehicles, in which it can be argued that humans need to be in a significant degree of control, in order to compensate for any system failures (using a full range of urgency). The type of intervention we recommend then lives within the urgency continuum, and depends on the urgency and better understanding it. As each task has a certain time window to be executed, the elapsed time of an omitted or wrong executed task plays a major role in the choice of the adequate intervention of any assistant system. Thus in order to enact any of these solutions, a taxonomy of sorts should be built or generated that allows for some speculation and characterization of task performance by the system and the human. Because these can be done, and methods are widely known to accomplish it [5, 6] we focus instead on the interactions rather than the task properties.

2 Motivation

Highly automated systems, which are able to detect human errors, are also typically designed in a way that they immediately correct human errors. This approach dispossesses the human operator of his/her task immediately, independent of whether the human operator still has enough time, mental resources, and/or the ability to correct the error on their own. If such an error correction occurs often and the correction is relatively reliable, this may cause complacency effects in the human. The human may put a miscalibrated, high amount of trust in the corrective actions of the automation, therefore neglect his own tasks and consequently lose vigilance or situation awareness [7]. In general, negative effects of automation may be avoided by actively involving the human operator in the error correction process itself. Therefore, within the current design pattern, we suggest a directed, stepwise-escalated error correction approach to support the human operator based on his/her needs, and the urgency of an emerging or an already occurred error.

This pattern will also provide a means for the human-autonomy team to adapt tasks and actions as urgency for completing tasks increases. This implies the need for repeated adaptation, as urgent tasks are completed, become less urgent, are abandoned, or are considered obsolete. Examples related to the importance of urgency are seen in autonomous assistants for aviation and driving. If the system determines that a collision risk is high from its sensor data, the system can then determine deadlines for the various forms of intervention based on the interactions between models of the autonomy and models of human performance. For our current purposes, we make the assumption that the autonomy's deadlines are generally later than the human's. As a human deadline approaches, urgency is increased and the system managing the tasks can invoke actions intended to reduce the risk that the deadline will be missed. These can include notifications, reprioritization of tasks, changes in methods for completing tasks, and task abandonment in the case of lower priority tasks (for example in terms of reward, or cost).

Important to this pattern is a definition of urgency, which we attempt to provide with regard to two task types. Consider a model of each available agent (human and/or machine) capable of performing tasks within a work process. For each task, the predicted performance is a probability density function representing the probability that an

instance of a task type will require any particular time by the agent for completion. For each task instance in each agent’s queue, there is an associated task type, and a required completion time. Tasks can be decomposed into subtasks and methods, such as in the structure in the figure below, if desired (and these form composite tasks), but is not necessary. Urgency of an atomic task is the simple probability that the task will not be completed on time given the current resource allocation (in this case, which agent is assigned and their capacity and state, for example) (Fig. 2).

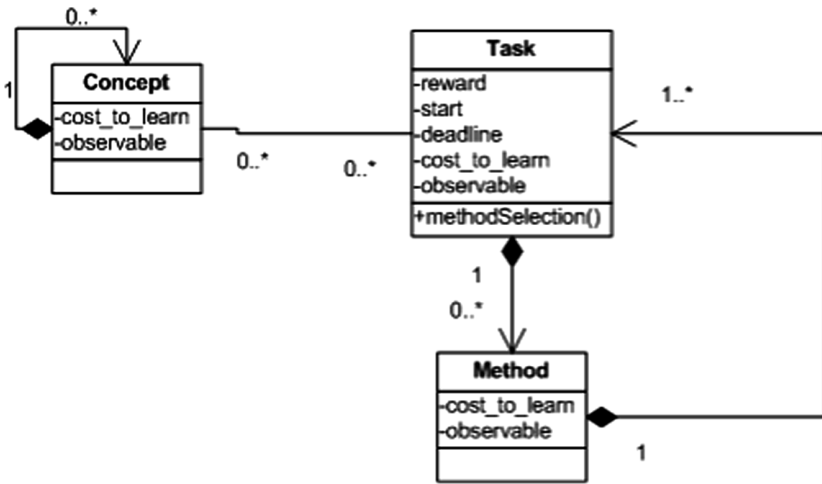


Fig. 2. Possible task structure.

Urgency for a composite task in contrast must be propagated based on a task decomposition pattern. Composite Tasks (one model of hierarchical WProc) are composed of methods, with an ‘or’ relationship. Any of the methods can be selected to complete the task. Methods are composed of tasks, all of which must be completed to complete the method. Therefore, our definitions are as follows:

- A task’s urgency is the probability that it will not be completed in time if it is atomic; else if
- A method’s urgency is the maximum urgency present among its tasks.
- A hierarchical/composite task’s urgency is the (median | mean | max | min | $E(x)$ | $CoV@R(r, x)$) urgency of the methods it is composed of. The exact method is up to the system designer.
- A networked WProc can be modelled using composite tasks for the purposes of urgency. There are at least 3 subtasks (receive input, perform task, and send output). “Perform task” is likely to be further decomposed to account for aspects that must periodically wait for input.

3 Applicability

Use this pattern when you want to mitigate complacency effects in a human operator, which might be caused if technical systems always attempt to perform an immediate and automated error correction.

Do not use this pattern for time critical support of human operators, where you need immediate function or task adoption by the technical system. This design pattern is more apt for tactical or strategic tasks where the human can contribute, rather than reactive tasks.

4 Structure

Figure 3 illustrates the collaboration between human and agents or automation within a work system, which enables a “step-by-step” error correction. It uses a graphical language, which is defined in [8]. On the left hand side both the human and the assistant system are workers. Workers know the given work objective. They are able to understand and pursue the work objective according to their abilities. The relation between both workers (human and assistant system) is a heterarchical one (blue connection); therefore within this cooperation schema there is no hierarchical order guiding the involved workers. Instead, each worker acts on its own initiative to pursue the overall mission goal. In this example case, the assistant system continuously monitors both the human and the work process, and supports the human step-by-step in the achievement of the overall mission goal. While pursuing the work objective both workers use the available tools. These tools are subordinate to the worker, as shown on the right hand side of the work process. These tools are in a delegation/supervisory control relationship to the worker (green connection). Workers receive tasks, instructions, or commands which they have to execute in order to achieve the given/delegated tasks. Within this work system we describe two distinct kind of agents, each which have differing purposes [9]:

- Purpose of the delegate agent: Control of conventional automation to reduce or remove human task loading.
- Purpose of the assistant system: Mitigation, i.e. prevention or correction of erroneous behavior of the human, in order to maximize safety but avoid complacency behaviors.

5 Participants

As depicted in Fig. 3 the participants are a human and at least one intelligent agent, the agent on the worker side, referred to as “assistant system.” The human is in charge of the achievement of the given mission objective. The assistant system lets the human accomplish his/her tasks as long as no errors (errors of omission, errors of commission) emerge. In case these errors occur, the assistant system chooses an adequate intervention strategy according to the urgency of the task, which has either to be

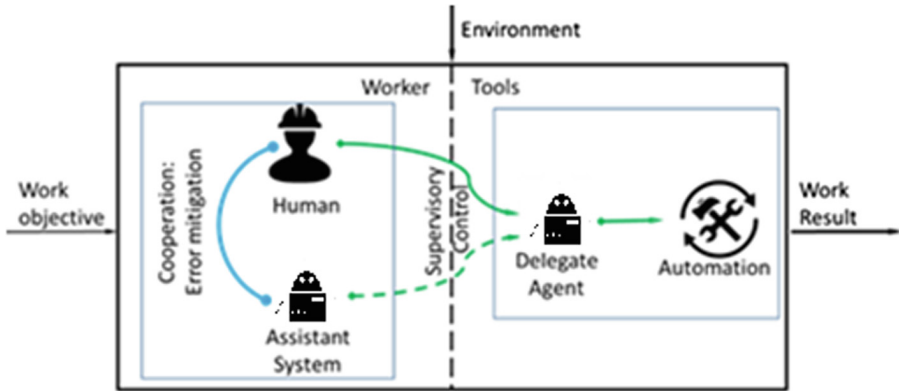


Fig. 3. The high-level elements for a design pattern, “increasing urgency/step-by-step error correction” (Color figure online)

accomplished on time or corrected for an error. A working agreement (see section on this pattern) between the human and the assistant can be agreed and in place. This allows the human to establish the parameters of how the assistant will behave as urgency increases, and informs and constrains the actions of the assistant to ensure the human’s mental model matches that of the agent when actions are needed [see for example, 10].

The “delegate agent” that appears on the tool side of Fig. 3 is optional. This agent is able to accomplish given tasks by use of available automation. If this agent is not available, the interaction of the human or the assistant system is on the level of commands for the available conventional automation.

6 Collaborations

The fundamental requirement for this pattern is that there be at least one participant agent that is aware of the urgency and priority of the tasks. The participant should have an agreement with any humans or other agents concerning the rules under which actions may be taken. Such agreement can be built into the system, or more flexibly created using a working agreement design pattern [please see 11, for required details].

A working agreement is a task-centric, shared understanding of how task performance is to be split and shared between partners. These styles of agreement can be found in air traffic control, for example, in splitting up airspace responsibilities [e.g., 12]. Working agreements between humans and automation should be accompanied with several benefits to the each agent as well as the system overall – first, the development of the agreements helps articulate the tasks and methods required to perform them for both the agent and the system (a step not always taken in system design). Second, an agreement helps in understanding how these tasks should be allocated effectively and allows for evaluation (agreement A versus B). Third, the definitions of agreements allows their codification into system- and human-understandable display. In other

words, agents in the system get clarification on what other agents are doing and supposed to do given a set of conditions [11]. Usually the human does not have this level of awareness in a system, leading to mental model mismatching.

7 Consequences and Justification

This pattern mainly affects the cooperation between the human and the assistant system. The assistant system in general behaves like a restrained human teammate. Within normal situations, the assistant system is no more than a silent observer. Only in situations which require an action of the assistant system to prevent a degradation of the overall mission performance, the assistant system becomes active with a situation adequate intervention, falling within the working agreement structure. This restrained behaviour of the assistant system will leave the human in charge of task accomplishment, as long he/she is able to do his/her task on their own according to estimations and current projections. In these times where it is necessary for the human to take actions (e.g. recognition of a effecting change in situation, necessary execution in tasks) the assistant system makes an appearance, by giving alerts, hints, or messages without wresting the human from his task. The human will be kept in the loop and supported as long as the human has enough time, resources and capabilities to solve the situation on his/her own.

In many ways, this positive benefit harkens to “lockout” or constraint methods of processing, in which certain actions that are harmful are literally prevented by manipulating the interaction capabilities (or removing a capability altogether under certain circumstances). An example is the “grey out” of action buttons on an interface; not only does this prevent the user from making an inappropriate response, but it can also communicate that the system believe it is inappropriate. Similarly, other changes in design and lockouts – such as those used to prevent sudden unintended gear changes in vehicles, and those made to physical equipment (such as changing the fittings on operating room equipment to avoid connecting the wrong gas tanks to patients) provide major safety improvements that greatly reduce the burden on the human operator to “avoid error.” These system-driven error reduction methods come highly recommended from other engineering domains and are at the heart of major theoretical advances in human error mitigation [13, 14].

As discussed, the difficulties here lie in determining what those actions are during system design and not in hindsight after an accident or devastating error is committed. Presumably, we can account for a large portion of both, but never all of either type. This leads to conditions when the human may need access and the design blocks it; or times when the design fails to block an action that leads to mistakes.

Another possible downside might be, that a human can adapt to the restrained behavior of the assistant system – in other words complacency. This means the human could wait until no more time is available to do the task on his own, when the assistant system would then stand in by a full task adoption from the human.

8 Implementation

For the realization of an assistant system, which provides a stepwise increasing intervention policy, the assistant system has to have the following capabilities [7]:

- Monitoring of the environment and detection and analysis of danger
- Monitoring of the human and interpretation of the observed data with respect to the human's cognitive state(s)
- Planning and scheduling of interventions
- Execution of interventions, i.e. of the actual interaction with the human

The interventions of the assistant system can be supportive but reserved. The overall goal is, to keep the human in the loop as long as possible and responsible for task accomplishment. To enable this requirement the assistant system shall express the following desired behavior:

- The human shall be given as much time as possible to find own solution
- Interventions shall provide input that helps with the current problem (but may not solve it as optimally as a human expert)
- Dangerous situations shall be resolved before fatal/critical damage is inflicted
- The input given by an intervention shall not exceed the current problem.

In order to identify the emerging conflict situation, the urgency, and selection of the adequate intervention strategy, the assistant system has to continuously:

- Determine if the current situation is dangerous and, if so, at what time damage (a violated threshold of certain performance parameters) will be inflicted. A situation is dangerous if the further development, without intervention by the human or the assistant system will lead to damage (e.g. degradation of the overall work result).
- Determine what the human should do to resolve the dangerous situation. The resolution typically consists of giving a certain command (sequence) either to an existing delegate agent or to conventional automation.
- Estimate the current cognitive state of the human. The cognitive state includes mental resources such as situation awareness, vigilance, workload and focus of attention. It also includes the state of information processing, i.e. the current task(s) and the associated cognitive processes. These estimates should be based on a model of the human's information processing but could be informed by real-time inputs and measures.
- Compute the transitions of the human's cognitive states leading from the current situation to the resolution of the dangerous situation, and identify the conditions for each transition: What steps will the human's mind have to go through to effect the desired action, beginning with its current state? These steps and estimates of their duration (with buffers and worst-case assumptions) should be based on a model of the human's information processing.
- Arrange these computed mental steps along a timeline, beginning with the earliest one. Arrange them in a way that the final step (the desired action) takes place

immediately before the moment of damage, i.e. barely in time. The position of the left end, i.e. the starting point of the sequence, will then determine the point in time at which the pilot must begin working on the problem in order to prevent damage in the worst predicted case

- Determine whether the starting point of the sequence is in the future, or not?
 - (Yes): There is still time left for the human to find own solutions. The system should do nothing.
 - (No): The human should have reacted by now. Intervene by enforcing the current transition, i.e. the first step, using any available means.

In Fig. 4, we provide an example which shows how the assistant system derives necessary interventions. Within this example, a human operator has to enter new commands to his own aircraft to avoid a collision with a foreign aircraft. So the task for the operator is to enter the right evasion commands (2) to avoid the collision. This has to be happened latest immediate before time (1), which is the last chance to avoid the damage. In fact, the human operator is actually analyzing his tactical map (3). By monitoring the human the assistant system could detect, that the human did not yet detect the foreign aircraft. The assistant system determines that the sequence of detection, information processing and action, leading from the actual task of the operator to the desired action steps (4, 5) will likely not be completed in time (6a). Therefore, the assistant system intervenes (6b): It enforces the transition from the human operator's current mental state to the next state (detection of a relevant change in the tactical environment) by alerting the human operator about an incoming other aircraft

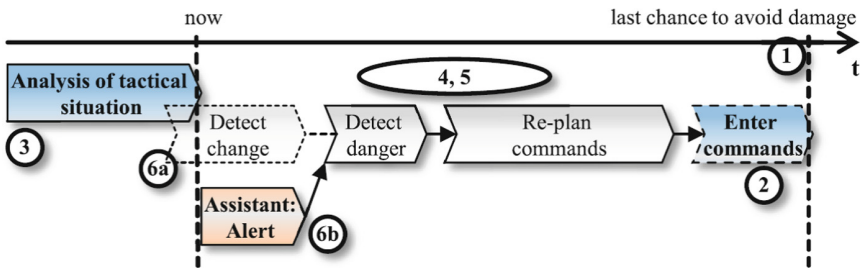


Fig. 4. Example for planning and scheduling of interventions by the assistant system

For each possible dangerous situation, the assistant system has to have a repertoire of different intervention possibilities, e.g.:

- neutral alerts without hints to the emerging error situation,
- directed alert towards the emerging problem,
- messages and suggestions to give hints to the human how to solve the situation,
- proposals to adopt part-tasks to support the human in task accomplishment,
- complete task adaptations in temporal critical situations

One possible consideration is that behavior may be different when urgency is increasing and when it is decreasing. The working agreement with the task manager should specify what these behaviors should be [15]. Below is an example table of behaviors that may explain this better than a formal definition (Table 1).

Table 1. Sample urgency decision table.

| Notional urgency range | When urgency increasing since last time step... | When urgency decreasing since last time step... |
|------------------------|--|--|
| 0–24 | No change from standard working agreement | Treat as having indicated priority. Remove requests to abandon or messages and failure and replace with explanation dialog. If activity on method has not been initiated by human and another method requiring more human action is available that will keep task within this urgency range, then change methods |
| 25–49 | If human activity has not been initiated on this method and another method is available that reduces urgency, then select new method | Remove highlights in lists. Remove requests to abandon or messages and failure and replace with explanation dialog |
| 50–74 | Use alert highlighting in lists. Treat as having 50% higher priority in sorts | If a request to abandon task or a failure message has been presented, remove message. Provide dialog explaining reason for removal |
| 75–99 | If there are no task dependent, then request approval to abandon task | If a failure message presented, remove message and provide dialog explaining removal |
| 100 | Failure message to human. Request extension on deadline or abandonment of task | Not possible |

9 Examples and Known Uses

This pattern has been applied to the domain of unmanned air reconnaissance conducted by a single human pilot in a ground control station.

The work objective of the single human pilot was to gain reconnaissance information on certain objects (buildings, persons, vehicles) in a hostile area. The required information could be obtained by using the sensors attached to an unmanned aircraft. These sensors provided video and imaging data to the human pilot. Beside the task of gathering and evaluation of sensor data to gain the required information, the pilot has

furthermore to manage the flight of the unmanned aircraft. The reconnaissance targets were given to the pilot beforehand, but could have changed during the mission. The execution was also constrained by airspace regulations (boundaries and corridors), threats (possible unexpected hostile air defenses), and resource limitations (fuel). As it was a single pilot station, the pilot had to carry out the tasks of flight management, sensor management and interpretation of sensor data in parallel. Therefore, the pilot was supported by an assistant system according to the described design pattern.

Within this use-case, the assistant system had to prevent, among others, the following effects of erroneous behaviour of the human pilot:

- Violation of airspace regulations by the unmanned aircraft
- Loss of unmanned aircraft by exhaustion of fuel reserves during flight
- Loss of unmanned aircraft by entry into the threat radius of hostile air defense sites
- Ineffective reconnaissance (inadequate fulfilment of the mission objective)

To avoid these effects, the assistant system was allowed to intervene. It was integrated into the control station's systems and had direct access to the pilot's GUI. Depending on the information processing step of the human determined by the assistant system, the assistant system was able to display general alerts and iconic or textual messages, highlight certain screen elements, direct the pilot's attention to other screens, or override commands if necessary. The assistant system gathered all necessary information to plan, schedule and execute an intervention from the subsystems. For a more detailed description of the implementation of the required functionality, please refer to [9].

An example of an escalating sequence of interventions of the assistant system in response to potential emerging violations is shown in Fig. 5 below.

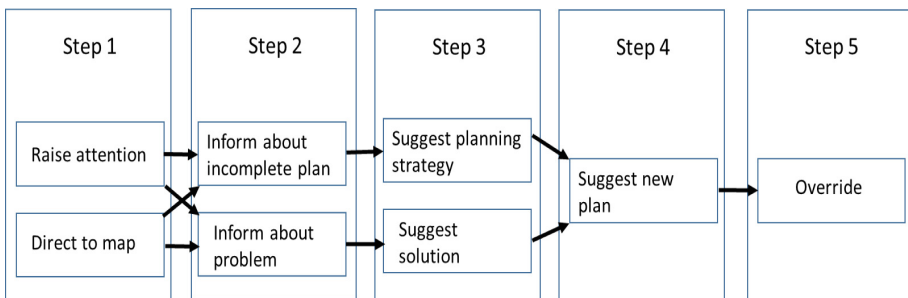


Fig. 5. Example for planning and scheduling of interventions by the assistant system

The following pictures (Fig. 6) illustrate one realization of the cooperation between the human operator and the assistant system by applying the stepwise escalating intervention sequence.

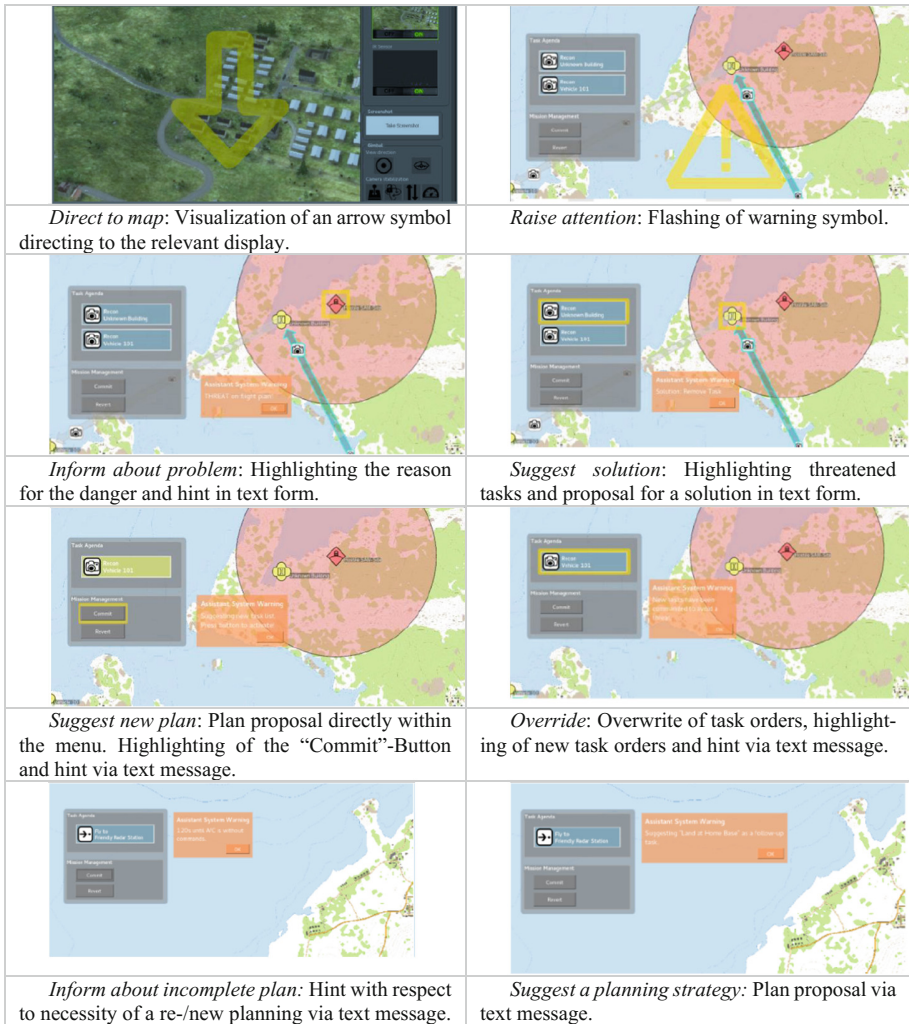


Fig. 6. Applying the stepwise escalating intervention sequence.

10 Related Patterns

This pattern can make use of working agreements [10, 11] to establish the rules by which urgency will be addressed. The formalization takes features from a pattern that might be titled tasks with deadlines and rewards. In that pattern, rewards are only received in full if the task is completed prior to its deadline.

This pattern has a complex interaction with the pattern human takes control upon autonomy failure. That pattern requires that a method requiring human attention be selected for a task that formerly was being performed by the autonomy. This can cause an immediate increase in urgency. This may be how the user is notified of the need to

take control (e.g., a new method is selected when the autonomy fails, this method requires urgent attention, so the task manager tries to reduce urgency by going to the autonomy). To avoid infinite loops, the methods allowing action by autonomy need to be marked as unavailable through some means.

References

1. Mosier, K., Skitka, L., Heers, S., Burdick, M.: Automation bias: decision making and performance in high-tech cockpits. *Int. J. Aviat. Psychol.* **8**(1), 47–63 (1998)
2. Parasuraman, R., Manzey, D.: Complacency and bias in human use of automation: an attentional integration. *Hum. Factors* **52**(3), 381–410 (2010)
3. Wickens, C., Hollands, J., Banbury, S., Parasuraman, R.: *Engineering Psychology and Human Performance*, 4th edn. Pearson, Upper Saddle River (2013)
4. Trapsilawati, F., Qu, X., Wickens, C., Chen, C.: Human factors assessment of conflict resolution aid reliability and time pressure in future air traffic control. *Ergonomics* **58**(6), 897–908 (2015)
5. Crandall, B., Klein, G., Hoffman, R.: *Working Minds: A Practitioner’s Guide to Cognitive Task Analysis*. MIT Press, Cambridge (2006)
6. Endsley, M., Jones, D.: *Designing for Situation Awareness: An Approach to Human-Centered Design*, 2nd edn. CRC Press, New York (2012)
7. Parasuraman, R., Molloy, R., Singh, I.: Performance consequences of automation-induced ‘complacency’. *Int. J. Aviat. Psychol.* **3**, 1–23 (2009)
8. Schulte, A., Donath, D., Lange, D.S.: Design patterns for human-cognitive agent teaming. In: Harris, D. (ed.) EPCE 2016. LNCS (LNAI), vol. 9736, pp. 231–243. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_24
9. Theiβing, N., Schulte, A.: Designing a support system to mitigate pilot error while minimizing out-of-the-loop-effects. In: Harris, D. (ed.) EPCE 2016. LNCS (LNAI), vol. 9736, pp. 439–451. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_43
10. de Greef, T., Arciszewski, H., Neerincx, M.: Adaptive automation based on an object-oriented task model: Implementation and evaluation in a realistic C2 environment. *J. Cogn. Eng. Decis. Mak.* **4**(2), 152–182 (2010)
11. Gutzwiller, R.S., Espinosa, S.H., Kenny, C., Lange, D.S.: A design pattern for working agreements in human-autonomy teaming. In: Cassenti, D.N. (ed.) AHFE 2017. AISC, vol. 591, pp. 12–24. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60591-3_2
12. US Department of Transportation Federal Aviation Administration, Air Traffic Organization Policy: Section 3. Letters of Agreement (2010)
13. 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)
14. Reason, J.: *Human Error*. Cambridge University Press, Cambridge (1990)
15. Lange, D.S., Gutzwiller, R.S.: Human-autonomy teaming patterns in the command and control of teams of autonomous systems. In: Harris, D. (ed.) EPCE 2016. LNCS (LNAI), vol. 9736, pp. 179–188. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_19



A Multidimensional Workload Assessment Method for Power Grid Dispatcher

Bingbing Song², Zhen Wang¹(✉), Yanyu Lu¹, Xiaobi Teng²,
Xinyi Chen², Yi Zhou², Hai Ye², and Shan Fu¹

¹ School of Electronics, Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
b2wz@sjtu.edu.cn

² East China Branch of State Grid Corporation of China,
Shanghai 200120, China

Abstract. Dispatcher's error is an important factor affecting the safe operation of power system. One of the main causes of human error is inappropriate workload. Due to the particularity of the power dispatching work process, existing workload measures are not ideal for power dispatcher. According to the human information processing model, combined with the actual work of dispatchers, this article proposed a novel method for dispatcher workload assessment. It considered dispatcher's workload from four dimensions: information perception, speech output, action output and attention. Video, audio and physiological monitor were deployed to acquire descriptive features. The frequency of incoming calls was extracted to describe information perception. Short-term energy and spectral entropy of the speech signal were extracted to describe speech output. Body movement speed was extracted to describe action output and heart rate was used to describe attention. The method was applied to an experiment in the dispatcher training simulator involving qualified power dispatchers. The experimental results showed that the proposed method was applicable and it can effectively reflect changes of dispatcher's workload during troubleshooting tasks.

Keywords: Power dispatcher · Workload · Multidimensional assessment

1 Introduction

With the progress of science and technology, the reliability of automations in complex systems has been greatly improved. It has been accepted in several areas (e.g. civil flight, air traffic control, nuclear plants and road traffic, etc.) that human factors have gradually become the primary threats to safety. Above 70% of accidents are relevant to human errors [1].

A power dispatching room is a typical human-in-the-loop complex system. In the power dispatching room, dispatchers handle with customers' requirements and release commands to power plants via communication system so as to make sure the energy supply meets customers' demands. Meanwhile they also need to monitor the status of the power grid so that the energy won't damage the plants and systems. As the

dispatchers do not directly operate the electricity production equipment, their cognition and decision on the grid status play a key role in safe operation of the power system [2].

The relationship between safety risk and workload is like a u-shape curve [3]. Too low or too high of the workload will both increase the risk to the system. When workload is too low, it is insufficient to maintain operator's situation awareness, and would decrease the speed and accuracy of operator's reaction. That would be dangerous especially in emergency situations. When the workload is too high, it might exceed operator's capability and would also degrade the quality of their performance.

Workload is an abstract concept which cannot be measured directly. According to previous studies, various techniques have been proposed to reflect workload. They could be classified into three broad categories [4]: (1) Subjective ratings, such as NASA-TLX, SWAT, Bedford and etc. these methods tend to quantitative operator's experienced workload through a set of elaborately designed questionnaires. (2) Performance measures, such as primary task performance and secondary task performance. These methods are based on the supply-demand relationship of mental resource. (3) Physiological measures, such as ECG, EOG, EEG and etc. these measures are based on the adjustment mechanism of autonomic nervous system.

However, the special work environment of power dispatcher restricts the application of the existing workload evaluation method. Firstly, power dispatcher's work is a continuous operation with long duration. Each shift has to work 8 h continuously (3 shifts in 24 h). Therefore, continuous monitoring of dispatcher's workload is essential for promptly detection of potential risk. Subjective rating techniques are post evaluation methods. Evaluation result can only be given after the work or every once in a while during the work. This would lead to low temporal resolution and would interfere with dispatcher's work if it was carried out during the work. Secondly, the role of power dispatcher is more like a commander than an operator. They release orders to other departments in the grid system and the order would be achieved by field personnel. The outcome of dispatcher's decision would return after an uncertain delay. Therefore, it is hard to use performance measures to evaluate dispatcher's present workload. Thirdly, the main function of power dispatcher is monitoring and decision making. The mental workload is much higher than physical workload. It should be verified that whether the physiological parameters approved in other conditions are still effective in the power dispatching room.

Currently, most of the conventional power dispatcher workload assessment methods are still based on time line analysis i.e. ratio of task occupied time to available time. For example, dispatcher's workload could be calculated as:

$$workload = (n * t + T_s) / (N * t + T_s) \quad (1)$$

where n is the total items in all the operation orders released by dispatcher, t is the average time to release one order item, T_s is the time occupied by secondary tasks, N is the maximum number of items the dispatcher can release (releasing operation orders to the control room operators is the primary task of a power system dispatcher, other tasks such as filling in the dispatch log, creating the order draft, verifying the operation order, approving the maintenance application, etc. are all secondary tasks).

Although task amount or time occupancy seem to reflect dispatcher's work intensity, it is not an effective assessment of dispatcher's workload. Sometimes, the

operator apparently has many operation activities, however, these activities might not take much mental resources. In fact, dispatcher's workload depends on the occupancy of his mental resource.

In order to develop an applicable method which can provide continuous and effective assessment of power dispatcher's workload, this study proposed a novel workload assessment model which both considering the characteristics of power dispatching work and the cognitive theory. According to this conceptual model, data acquisition, data processing and feature extraction methods were also introduced in this paper. Furthermore, we conduct an experiment in the power dispatching room so as to apply the proposed methods and test its effectiveness.

2 Method

2.1 Conceptual Framework

According to the human information processing model [5], there are several stages when human performing a task: (1) Short term sensory store (STSS). During which stage human acquire the outside information. (2) Perception. During which stage a person make a quick understanding of the information's apparent meaning. Combining working memory and long-term memory, human can make a further understanding of the situation. (3) Response selection. During which stage human weighs the pros and cons to make a decision. (4) Response execution. During which stage the brain control muscles to carry out the decision.

Besides the above four stages, there are two essential elements in the information processing model: feedback and attention. When the execution of response changes the environment and brings new pattern of information, feedback should be considered. Attention is the inherit capacity of human. It represents the occupancy of mental resource.

Note that there are not explicit boundaries between some stages in the information processing model. Those stages are carrying out swiftly in mind (e.g. STSS, perception and response selection, etc.) and cannot be measured separately. Furthermore, consider the separation of dispatching room and operation site and the delay of the order outcome, we do not consider the changes of dispatcher's response to the environment for the time. Therefore, we simplify the information processing model as Fig. 1. The simplified model includes two stages (cognition stage and response stage) and the attention elements.

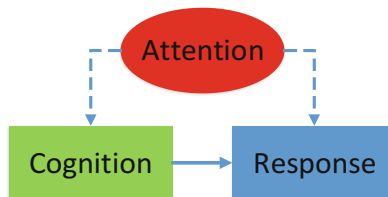


Fig. 1. A simplified human information processing model.

Combine the information processing model with power dispatcher's work. In the cognition stage, information primarily come from incoming calls. (Of course there are also visual information. However, we have not found an appropriate way to measure the amount of visual information acquired by a dispatcher). In the response stage, on one hand dispatchers respond in the form of action and movement such as keyboard input, checking for material, reaching for the phone and etc.; on the other hand, dispatchers respond in the form of speech such as giving operation orders or asking for situation. Both the cognition stage and the response stage cause dispatcher's attention. The attention or mental resource occupancy cannot be measured directly. In this study, we use physiological reactions to reflect dispatcher's attention.

Therefore, we build the following dispatcher workload model (Fig. 2). Workload can be reflected from four dimensions: information perception, action output, speech output and attention. Each dimension should be continuous measured. For example, information perception and speech output can be measured with audio acquisition. action can be monitored by video surveillance. Attention can be measured by physiological measures.

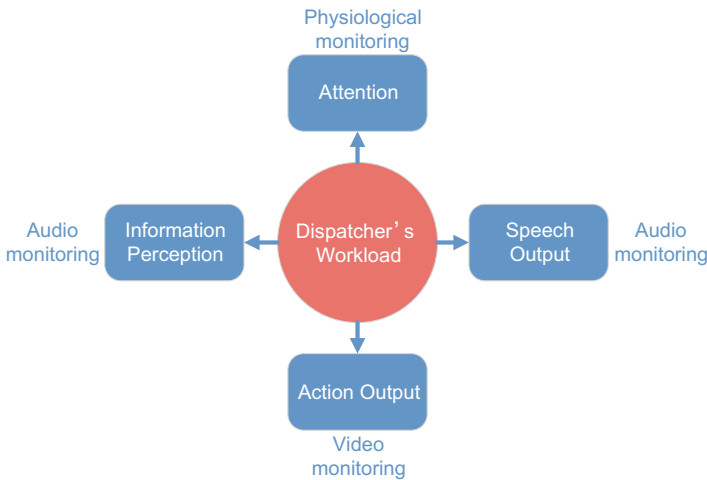


Fig. 2. The conceptual model of power dispatcher's workload.

In order to provide quantitative workload assessment result, each dimensions need to be represented by quantitative features. The following part of this section introduce our data processing and feature extraction methods.

2.2 Feature Extraction

Information Perception. In information perception dimension, we use “the frequency of incoming calls” to reflect task demand imposed on power dispatcher during the cognition stage.

In order to obtain the frequency of incoming calls, some audio processing technologies are required. In consideration of the frequency selection characteristic of band pass filter i.e. it can pass frequencies within a certain range and reject frequencies outside that range, we decide to use the Type I Chebyshev filter to separate ringtone signal from audio data [6]. The Amplitude-frequency relationship of a type I Chebyshev filter is as follows:

$$|H_n(j\omega)| = \frac{1}{\sqrt{1 + \epsilon^2 T_n^2(\frac{\omega}{\omega_0})}} \tag{2}$$

where ϵ is the ripple factor, ω_0 is the cutoff frequency and T_n is a Chebyshev polynomial of the n^{th} order. The passband exhibits equiripple behavior, with the ripple determined by the ripple factor ϵ .

In order to design a desired band pass filter, the frequency characteristics of the ringtone signal in real dispatching room should be studied. In addition, the lower passband frequency, the higher passband frequency, the lower stopband frequency, the higher stopband frequency, the passband ripple and the stopband ripple should be set up carefully.

After ringtone signal has been separated from the audio data, the frequency of incoming calls can be easily obtained by counting the number of ringtones in particular time intervals.

Speech Output. In the speech output dimension, we use “short-term average energy” [7] and “spectral entropy” [8] to reflect power dispatcher’s work intensity during voice communication.

First of all, a band pass filter is also necessary in this dimension. By analyzing the frequency characteristics of real dispatchers’ voice, a band pass filter should be designed to separate voice signal from the audio data. Meanwhile, other sound such as ringtones and noises should be inhibited.

When voice signal is extracted from the original audio data, short-term average energy of the voice signal can be calculated by equation below. This feature describes dispatcher’s volume changes over time.

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \tag{3}$$

where E_n represents the short-term average energy of the n th frame, x_n is the n th frame of the voice signal segmented by a short-term window, N is the length of the short-term window.

Entropy is the measurement of uncertainty. Generally speaking, the more uncertain a signal, the more information it contains. In the voice signal processing area, spectral entropy can be used to describe the complexity of the speech. Spectral entropy can be calculated by the following steps:

- Segment the original voice signal $x(n)$ into frames
- For the n th frame $x_n(m)$, perform FFT to obtain $X(f)$
- The power spectral density is computed by $\text{PSD} = |X(f)|^2$

- Normalize the PSD so that it can be viewed as a probability density function.

$$p_i(f) = \frac{PSD_i(f)}{\sum_{i=0}^{N/2} PSD_i(f)} \quad (4)$$

where N is the length of FFT.

- Finally, spectral entropy can be calculated by:

$$H_i = - \sum_{k=0}^{N/2} p_i(k) \log p_i(k) \quad (5)$$

Action Output. In the action output dimension, “body movement speed” is used to describe dispatcher’s work intensity.

Image processing techniques are applied to compute body movement speed. The processing procedure is illustrated in Fig. 3. There are two major tasks in this procedure: motion detection [9] and skin detection [10].

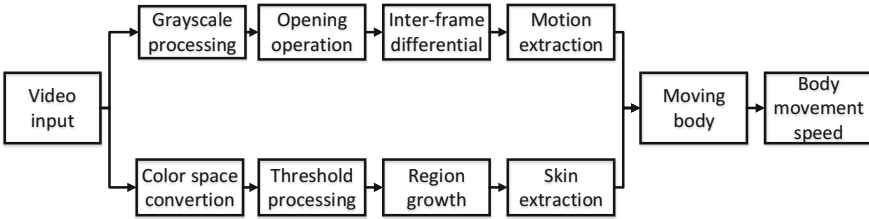


Fig. 3. The image processing procedure to obtain dispatcher’s body movement speed.

Motion detection takes 4 steps. (1) color image is converted to grayscale image. (2) “Opening Operation” is performed to reduce noise and filling holes in the image. (3) inter-frame difference is calculated so as to detect the moving part in the scene.

$$D_n(x, y) = |f_n(x, y) - f_{n-1}(x, y)| \quad (6)$$

where $f_n(x, y)$ is n^{th} frame of the video, $f_{n-1}(x, y)$ is the $(n-1)^{\text{th}}$ frame of the video. $D_n(x, y)$ is the difference of adjacent video frames. (4) By using threshold processing, the moving part can be explicitly distinguished from the background.

$$M_n(x, y) = \begin{cases} 255, & D_n(x, y) > T \\ 0, & \text{else} \end{cases} \quad (7)$$

where T is threshold. $M_n(x, y)$ is a black and white image in which moving part is white (255) and stationary background in black (0).

Skin detection takes 3 steps. (1) image is convert from RGB to HSV color space. (2) by using skin color threshold to extract body from image. This can roughly set the

skin areas to white and other areas to black. (3) using region growing to obtain relatively integral block to represent body part such as face, hand and etc.

Take the intersection of motion detection result and body detection result. This would provide a relatively robust representation of moving body. By calculating the area of all the white blocks, the result could represent dispatcher's body movement speed.

Attention. In the attention dimension, we use heart rate to represent the occupancy of mental resource. According to previous studies, heart rate has been proved to be an effective indicator of mental workload in various fields [11]. When operator encounters with more challenger tasks or feels more stressful, his/her heart rate would usually increase significantly.

Moreover, heart rate is easy to measure. It does not have to be measured by bulky laboratory instrument. Nowadays, heart rate can be measured unobtrusively by tiny remote sensors. For example, the Photoplethysmography (PPG) technology detects heart pulse by sensing the slight changes in the color of skin caused by blood flow. The PPG sensor is often very small which can be integrated into a common watch or wristband. The most important is that the PPG technique can provide satisfactory measurement accuracy [12].

Integrated Assessment. The proposed method requires different types of measure data and they may come from different instruments. In order to associate all the data for integrated workload assessment, time synchronization is an inevitable problem. In this study, we use sample timestamp to synchronize all the data (the premise is that all devices' internal clocks have been synchronized in advance).

After all the features have been synchronized, principal component analysis is used to examine the correlation between the features and combined them by using the following equation [13]:

$$W = \sum_{i=1}^n \beta_i z_i \quad (8)$$

where, W is the overall dispatcher workload assessment index. n is the total number of principal components. β_i is the percentage of variance explained by the i^{th} component. z_i is the component score of the i^{th} component.

3 Experiment

3.1 Participants and Apparatus

Ten qualified male power dispatchers participated in this experiment. Their age is between 28–32 years old and with the working experience of 3–6 years. The experiment was carried out in the Dispatcher Training Simulator (DTS) which is a system that can simulate the behavior of electrical network under various conditions. In this experiment, the instructor set up a series of adverse scenarios for them including different kinds of equipment malfunctions. The participants did not know these setting

in advance. They were supposed to quickly analyze the situation and give out reasonable solutions.

During the experiment, video data was collected by a HD wide-angle camera (Fig. 4). There is a microphone integrated in this camera, so that audio data can be collected concurrently. The video and audio data are stored offline in the camera's build-in memory card. We continuously collected dispatcher's heart rate with a heart rate watch (Mio Alpha; Physical Enterprises Inc., Vancouver, BC) which use PPG technology. Therefore, it was unnecessary for participants to wear sensors such as chest strap near their heart. Heart rate data is transmitted in real time to a receiving device via Bluetooth. In this study, the receiving device is a laptop with Bluetooth module, and we have developed a special software that can receive and store the heart rate data with sampling time-stamps.

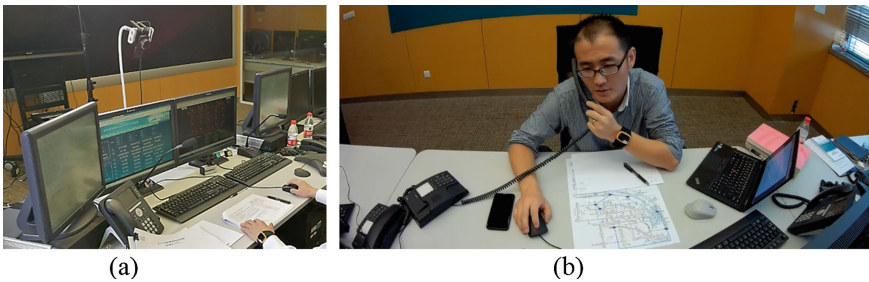


Fig. 4. Experimental setting in the DTS. (a) a HD wide-range camera was install above the monitor. (b) the dispatcher is wearing a heart rate watch on his left wrist

3.2 Experiment Procedure

The experiment was conducted in accordance with the following steps: (1) Before experiment started, experimenter installed the camera to ensure that it did not interfere with dispatcher's sight and fully cover the dispatcher's working area. (2) Experimenter synchronize the internal clocks of the camera and the heart rate receiving laptop. (3) After participant's entering to the DTS, experimenter helped him to wear the heart rate watch, started the heart rate collection function and test the validity of the data. (4) Experimenter started camera's recording function and the data receiving software in the laptop. (5) Experimenter left the DTS and started the experiment. Each trial lasted about one and a half hours. (6) After the experiment, the experimenter terminated the collection and exported the data.

3.3 Result

Information Perception. The waveform of a segment of ringtone and its FFT are illustrated in Fig. 5(a) and (b). According to the frequency characteristics of the ringtone, we design a band-pass filter. The amplitude-frequency curve of the band pass filter is shown in Fig. 5(c). Specifically, the lower passband frequency was 3300 Hz,

the higher passband frequency was 3450 Hz, the lower stopband frequency was 3200 Hz, the higher stopband frequency was 3500 Hz, the passband ripple was 1 dB and the stopband ripple was 5 dB.

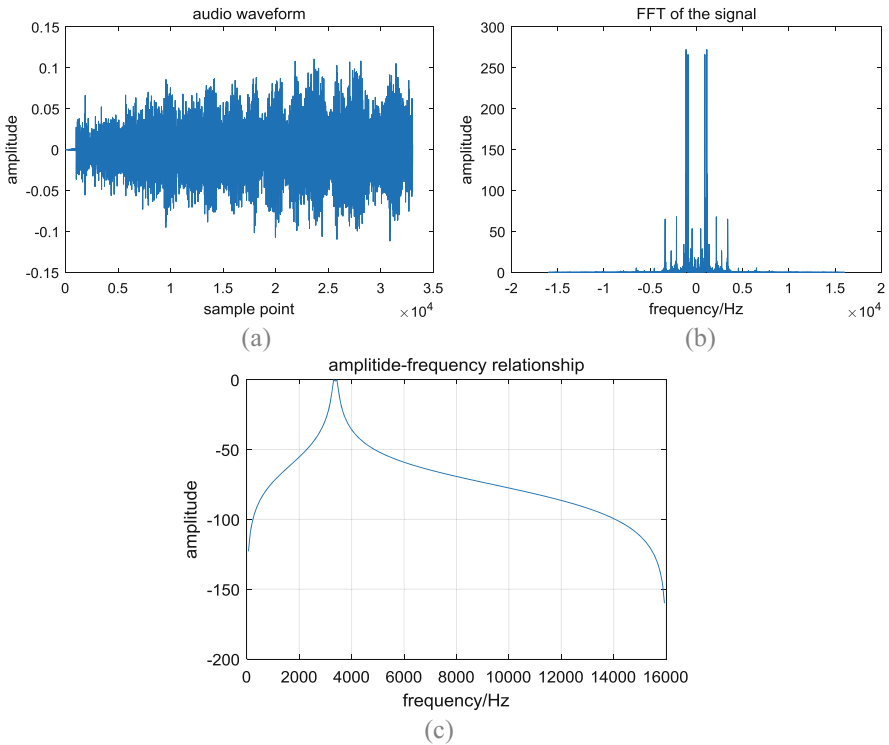


Fig. 5. Design of a band pass filter for ringtone extraction. (a) the waveform of the ringtone signal. (b) FFT of the ringtone signal, (c) The amplitude-frequency relationship of the band pass filter which is used to extract ringtone from audio signal.

A segment of audio signal containing both ringtone and speech was shown in Fig. 6 (a). After band-pass filtering, the result was shown in Fig. 6(b). As can be seen, this band pass filter can effectively separate ringtones from original audio.

Speech Output. Another band pass filter was designed to separate speech from audio. Its amplitude-frequency curve was illustrated in Fig. 7. Specifically, the lower passband frequency was 100 Hz, the higher passband frequency was 800 Hz, the lower stopband frequency was 50 Hz, the higher stopband frequency was 850 Hz, the passband ripple was 1 dB and the stopband ripple was 5 dB. As can be seen that after band pass filtering, ringtones in the original signal has been weakened and the speech became more significant.

After band pass filtering, the short-term energy and spectral entropy of the speech signal were calculated and plotted as illustrated in Fig. 8(a) and (b) respectively.

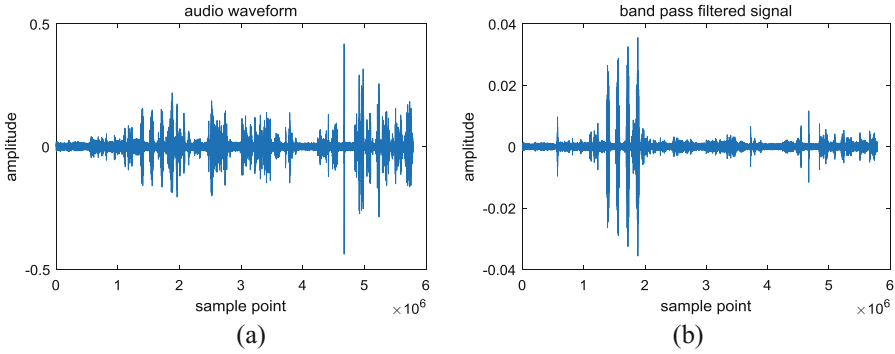


Fig. 6. Result of band pass filtering. (a) the waveform of a segment of audio signal containing both speech and ringtone. (b) Signal after band pass filtering.

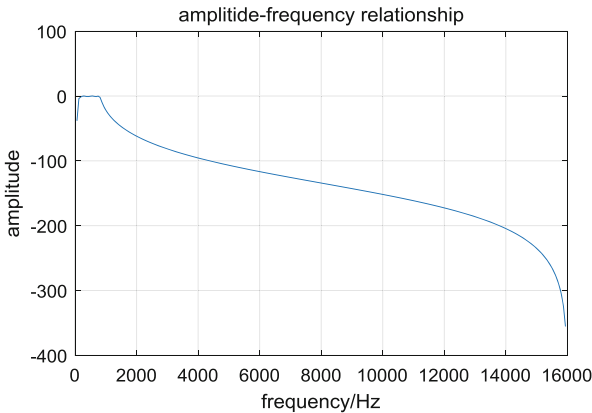


Fig. 7. The amplitude-frequency curve of the band pass filter which is used to extract speech from audio signal.

Action Output. By associating the results of skin detection and motion detection, the moving body part was detected as illustrated in Fig. 9.

The body movement speed can be calculated simply by counting the area of the white blocks in the result image. The changes of body movement speed over a period to time was illustrated in Fig. 10.

Attention. The original heart rate data was illustrated in Fig. 11(a). Due to body movements, sometimes the heart rate sensor would be poorly contacted. This would result in some outliers. After a simple threshold processing, we got the ideal heart rate data, as illustrated in Fig. 11(b).

Integrated Assessment. By performing PCA, five principal components were extracted from the features. The Eigen value and percentage of variance explained by each component was illustrated in Table 1.

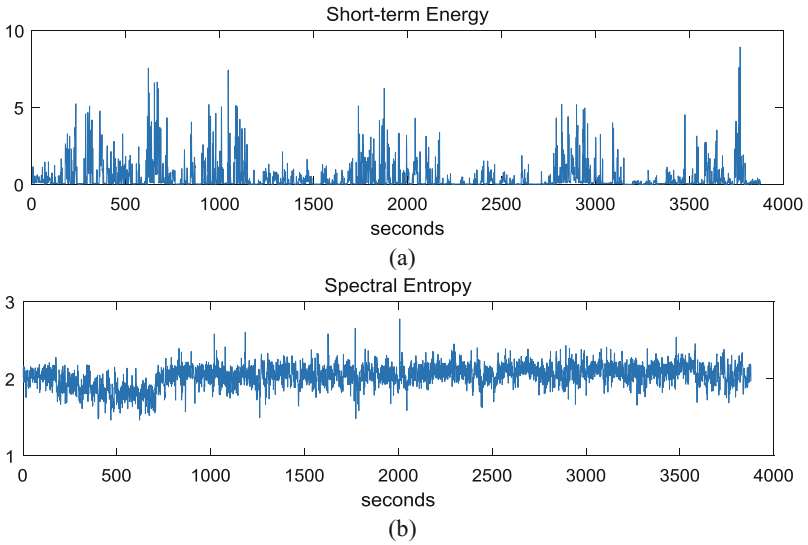


Fig. 8. Short-term energy and spectral entropy of a segment of speech signal. (a) Short-term energy. (b) Spectral entropy.

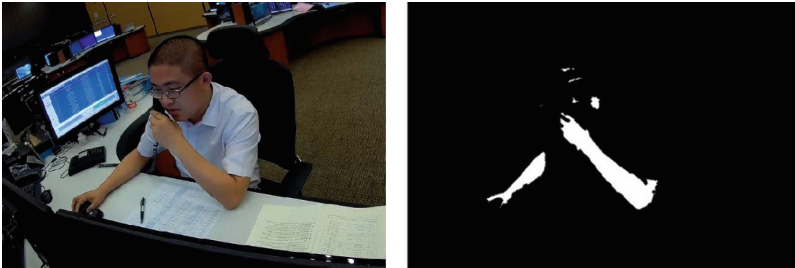


Fig. 9. Moving body has been extracted from video image.

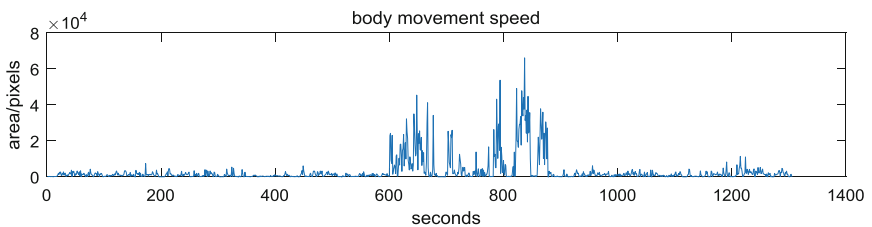


Fig. 10. The changes of body movement speed over time.

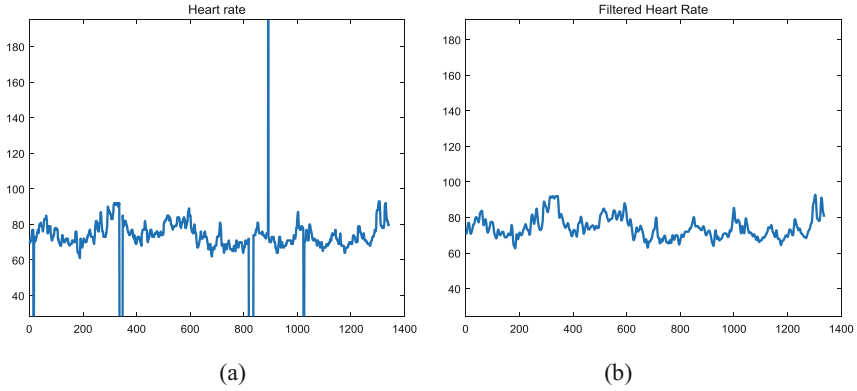


Fig. 11. The result of threshold processing of the heart rate data. (a) Original heart rate data. (b) Filtered heart rate.

Table 1. The result of principal component analysis.

| Component | Eigen value | Explained variance (%) | Accumulated explained variance (%) |
|-----------|-------------|------------------------|------------------------------------|
| z_1 | 0.0533 | 0.6902 | 0.6902 |
| z_2 | 0.0126 | 0.1637 | 0.8539 |
| z_3 | 0.0106 | 0.1379 | 0.9918 |
| z_4 | 0.0003 | 0.0042 | 0.9960 |
| z_5 | 0.0003 | 0.0040 | 1 |

According to Eq. 8, the overall workload of a power dispatcher can be calculated:

$$W = 0.6902z_1 + 0.1637z_2 + 0.1379z_3 + 0.0042z_4 + 0.0040z_5 \tag{9}$$

Figure 12 shows the changes of dispatcher’s overall workload over time.

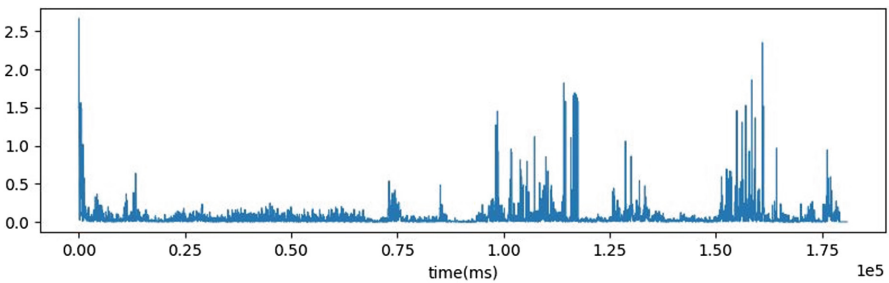


Fig. 12. The result of dispatcher’s overall workload.

4 Discussion

As can be seen from Fig. 12, the overall workload waveform significantly rises at 100000 ms and 150000 ms. By examining the original video and audio data, it was found that at those two moments there happened to be two incoming calls respectively. One is from dispatcher's superior asking for the condition of the malfunction, and the other is from field staff at the local transformer substation consulting troubleshooting proposal. For those two calls, dispatchers could not give definite replies because the cause of the malfunction had not been revealed yet. During these periods, the dispatcher needs to consider various factors to comprehensively evaluate the current situation and make appropriate judgement. It requires the brain to conduct more analysis and processing than usual. As a result, the resource occupancy should be higher. To a certain degree, this prove the validity of the proposed method.

There are still some limitations in this study. Firstly, in the information perception dimension, we did not consider information. This was because that we have not found an effective way to measure the amount of visual information perceived by dispatcher. Although eye tracking was a potential way [14], however, based on our actual inspection, we found that almost all the dispatchers in this study wear glasses. This limit the application of the existing glasses-type eye tracker. Considering the wide distribution of dispatcher's visual attention (they often need to monitor 5 to 6 screens at the same time), desktop eye tracker was also not applicable. Secondly, due to limited time, this study did not go further into the meaning of each principal component. Actually, it was very important for us to understand the mechanism of human factor risk in power dispatching. Thirdly, this study was based on data sample from relatively few trials. This is far from enough to reflect the general situation in the dispatching work. In future study, more experiments need to be carried out in both DTS and real dispatching room. Statistical analysis should be done to test the validity of the proposed method.

5 Conclusion

In this paper, we analyzed the human information processing model, and combined with the actual work of power grid dispatcher, put forward a novel dispatcher's workload evaluation method. The method considered that dispatcher's workload can be describe from four dimensions: information perception, speech output, action output and attention. In each dimension, the feature extraction method had been introduced in detail. Principal component analysis was used to combine all the features into an overall workload assessment. The proposed method was applied to an experiment involving certified power dispatchers. The experimental results proved the validity of this method to some extent. In future research, the sample size need to be increased. We will test the validity and reliability of the proposed method more rigorously with statistical analysis.

References

1. Pasquale, V.D., et al.: A simulator for human error probability analysis (SHERPA). *Reliab. Eng. Syst. Saf.* **139**, 17–32 (2015)
2. Prevost, M., et al.: Preventing human errors in power grid management systems through user-interface redesign. In: *IEEE International Conference on Systems, Man and Cybernetics*, pp. 626–631. IEEE (2007)
3. Grigoras, G., Bărbulescu, C.: Human errors monitoring in electrical transmission networks based on a partitioning algorithm. *Int. J. Electr. Power Energy Syst.* **49**, 128–136 (2013)
4. Cain, B.: *A Review of the Mental Workload Literature*. Defence Research and Development Canada, Toronto (2007)
5. Wickens, C.D., et al.: *Engineering Psychology and Human Performance*, 4th edn. Psychology Press, Hove (2012)
6. Podder, P., et al.: Design and implementation of Butterworth, Chebyshev-I and elliptic filter for speech signal analysis. *Int. J. Comput. Appl.* **98**(7), 12–18 (2014)
7. Lokhande, N.N., Nehe, D.N.S., Vikhe, P.S.: Voice activity detection algorithm for speech recognition applications. In: *IJCA Proceedings on International Conference in Computational Intelligence (ICCI)*, pp. 5–7 (2011)
8. Lee, W.-S., Roh, Y.-W., Kim, D.-J., Kim, J.-H., Hong, K.-S.: Speech emotion recognition using spectral entropy. In: Xiong, C., Liu, H., Huang, Y., Xiong, Y. (eds.) *ICIRA 2008. LNCS (LNAI)*, vol. 5315, pp. 45–54. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88518-4_6
9. Wang, Z., Wang, W.: LHMM-based gathering detection in video surveillance. In: *International Conference on Intelligent Computing*, pp. 213–216. IEEE (2010)
10. Raheja, J.L., Das, K., Chaudhary, A.: Fingertip detection: a fast method with natural hand. *Int. J. Embed. Syst. Comput. Eng.* **3**(2), 85–88 (2012)
11. Farmer, E., Brownson, A.: *Review of Workload Measurement, Analysis and Interpretation Methods*. European Organisation for the Safety of Air Navigation, Brussels (2003)
12. Wang, Z., Fu, S.: Evaluation of a strapless heart rate monitor during simulated flight tasks. *J. Occup. Environ. Hyg.* **13**(3), 185–192 (2016)
13. Ryu, K., Myung, R.: Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *Int. J. Ind. Ergon.* **35**(11), 991–1009 (2005)
14. Holmqvist, K., et al.: *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, Oxford (2011)



Task-Load Evaluation Method for Maintenance Personnel Based on the JACK Simulation

Ruishan Sun, Yuting Zhang^(✉), Zhen Liu, and Kang Li

Civil Aviation University of China, Tianjin, China
sunrsh@hotmail.com, zhangyt0715@outlook.com,
765609093@qq.com, kangkanggo@outlook.com

Abstract. The aim of the study is to establish a method to evaluate the task-load on maintenance personnel for civil aviation activities. Based on the characteristics of civil aircraft maintenance work and DORATASK Model from ICAO, a maintenance task-load index (TLI) is derived to measure maintenance personnel task-load. With the assistance of the JACK human factor analysis software, the method solves the shortcomings such as poor portability, subjectivity, and strict laboratory conditions in previous task-load evaluation methods. The proposed method provides a basis and a reference point for human resource allocation and occupational health damage prevention in civil aviation maintenance work.

Keywords: Aircraft maintenance · Task-load · Evaluation method · JACK

1 Introduction

The civil aviation transportation area has consistently grown in recent years in China, leading to a considerable expansion of fleet sizes. In conjunction with these changes, a shortage of professionals in China has become an increasingly prominent problem. Civil aviation personnel including pilots, air traffic controllers and maintenance technicians are trying to handle greater task-loads, and this is causing a strain, especially in the area of maintenance [11]. Researches show that 20%–30% of engine failure in air, 50% of delays, and 50% of flight cancellations are caused by human errors in maintenance areas [13]. In a study on human errors, errors were found to be caused by the volume of work accounts for a large proportion of errors, and this indicated that excessive task-loads may be one of the most important factors inducing or leading to flight accidents. The high task-load does not only affect the physical and mental health of the crew but also creates potential risks to civil aviation operation [2]. Therefore, with flight safety in mind, studying the evaluation criteria of civil aviation personnel task-loads becomes vital.

To solve the series of problems caused by excessive task-loads effectively, research on task-load evaluation methods has become a key subject in the field of aviation. Task-load is a part of workload. Till date, research on workload has mainly focused on controllers, pilots, and dispatchers. For example, in 1984, ICAO proposed evaluating the workload of controllers and dispatchers through the DORATASK theory [9]. Moreover, in 1992, Corwin evaluated pilot workload by using a subjective load assessment

method, which proved that the flight scores of high-load workers was higher than that of workers with lower workload under the same conditions [15]. In 1995, Tofukuji [14] discussed the workload of controllers based on airspace tension, checked whether the workload was acceptable, and then calculated the maximum actual traffic capacity. In 2000, Han [7] conducted several observational experiments on practical controls for air traffic approaches and area control airspaces in Guangzhou, China. He adopted the DORATASK method and analyzed the human factors that affected the capacity of sections. In 2012, Ren [12] analyzed the workload of the release dispatcher using single machine sorting to analyze the peak load and adopted a neighborhood variable to solve the assignment problem of dispatcher's seat assignment. In 2014, Sun proposed a control worker's workload measurement model that could represent the actual operation of Chinese air traffic controllers using a DORATASK load assessment method framework [16].

Based on the current research, the existing workload evaluation and measurement methods can be divided into two categories: subjective methods and objective methods. According to their characteristics and applicable scope, these methods can be divided into four categories: subjective evaluation methods, main task evaluation methods, auxiliary task evaluation methods, and physiological and biochemical index evaluation methods [18]. The application of these evaluation methods provides a basis for the evaluation and measurement of workloads, but they also have some limitations. The subjective evaluation method measures the workload according to the subjective feeling of the evaluation target. This evaluation result is accurate but it is too subjective. The main task and auxiliary task evaluation methods are highly dependent on the operational analysis ability of the researcher, so they are difficult to implement. The physiological and biochemical index evaluation method requires that the subjects wear various measurement instruments to observe physiological functions. Thus, the test cost is high, the study time consuming, the operation environment and experimental equipment are high demanded, and the laboratory requirements are strict.

2 Methodology

2.1 Research Framework

Based on an in-depth investigation and analysis of the maintenance work of civil aviation maintenance staff, this paper analyzes the impact of different workloads on the performance of the human body by using the JACK 8.4 engineering software to introduce how JACK is used in maintenance workload evaluations to provide a novel and portable system. The framework is as follow as Fig. 1 shows.

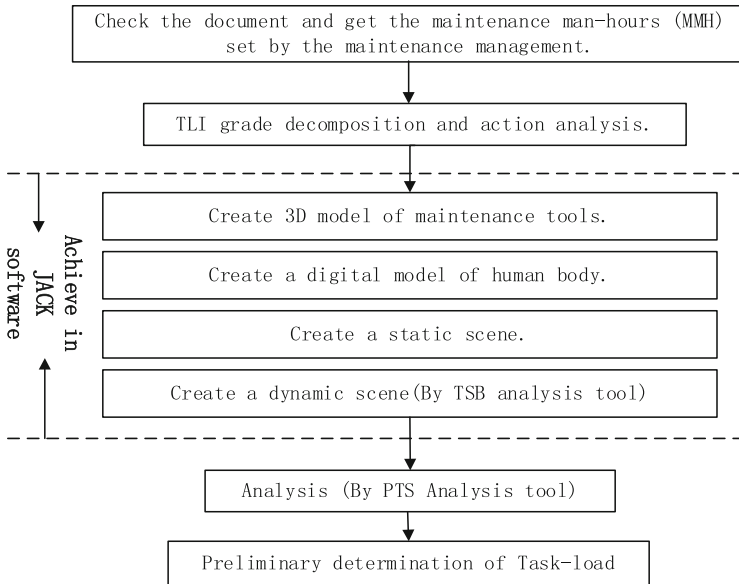


Fig. 1. The research framework of this paper

2.2 JACK Simulation Software Introduction

Jack is a human factors engineering analysis software from Siemens Industry Software (formerly UGS). Originally developed at the Center for Human Modeling and Simulation at the University of Pennsylvania in 1995, then commercialized by Siemens Industry Software more than 10 years. After more than ten years of research and improvement, the software has become a high-end simulation software integrating 3D simulation, digital human modeling and human performance analysis.

As a real-time visual simulation system, JACK can import CAD 3D model created by users and construct simulation environment. It can create a 3D human body model with biomechanical properties, assign tasks to digital human and obtain valuable information through the digital human behavior simulation analysis. This digital human model has been widely used in many fields of scientific research and engineering optimization design. The Jack Task Analysis Toolkit (TAT) enables researchers to evaluate human performance from an in-depth ergonomics perspective early in the product life-cycle, before designs are frozen and changes require costly rework. The toolkit can evaluate tasks using the Jack (male) and Jill (female) human models, without ever putting real workers at risk.

2.3 Maintenance Task-Load Evaluation Model

Civil maintenance work is characterized by high levels of time pressure, a complex working environment, standardized procedures and a substantial physical workload

[5, 6]. Based on the definition of working loads, this paper proposes a basic measurement called the “unit time task-load” as the maintenance task-load.

The “task-load” is one of the representation of workload. Workload is the quantity and quality of work tasks. Civil aviation engineering maintenance is a complicated and difficult work, so it is difficult to measure it directly. ICAO and CAAC used the DORA-TASK theory during the workload evaluation of air traffic controllers and dispatchers. First, they classified the work into specific categories. Then, through the observation and measurement of each sub-task in the task process, they used the “actual time consumed” to measure the workload [1, 8].

In this paper, the basic idea is to measure the task-load based on the rate of time occupancy. Using the above approach, we proposes the following task-load evaluation measurement model for maintenance personnel:

$$TLI = \frac{T_s}{T_a} \quad (1)$$

In Eq. (1), TLI is the task-load index; T_s is the simulated working time of a standard maintenance operation process through JACK; and T_a is available time provide from maintenance management of a maintenance task.

Here TLI are classified six grades:

Grade 1: $TLI \leq 0.2$, that the task-load is almost non-existent, and the available maintenance man-hours (MMH) needs to be re-planned;

Grade 2: $0.2 < TLI \leq 0.4$, that the task-load is low, and the currently scheduled MMH may be redundant for the actual maintenance task. The maintenance management needs to reduce the scheduled time of this task, in order to optimize the efficiency of maintenance;

Grade 3: $0.4 < TLI \leq 0.6$, that the task-load is relatively low, the MMH can be adjusted according to the actual situation;

Grade 4: $0.6 < TLI \leq 0.8$, that the task-load is good, the current MMH is basically the same as that required in the actual work, do not need to modify;

Grade 5: $0.8 < TLI \leq 1$, that the task-load is high, the maintenance management needs to increase the time required for maintenance, for reducing the task-load of maintenance personnel;

Grade 6: $TLI > 1$, that the MMH is not reasonable, this maintenance task is impossible to be achieved and must be reformulated.

2.4 Maintenance Task-Load Evaluation Based on JACK Simulation

On account of the specific nature of the maintenance work, the existing task-load evaluation method has a series of disadvantages in both its applicability and evaluation methods. Finding a task-load evaluation technique, which is characterized by high operational and low process invasiveness, and is also customized for the characteristics of civil aviation maintenance, is essential. In this paper, we put forward a method to evaluate the task-load using the Task Simulation Builder (TSB) system and the Task Analysis Toolkit (TAT) in the JACK 8.4 software.

Civil aviation maintenance work involves a large number of project types, and the various maintenance programs differ to a large extent. To prevent the formation of maintenance errors and ensure the quality of maintenance, the maintenance management has standard operating procedures for each maintenance project. Strict rules about the order of the maintenance and the size of the tools exist [10]. These rules provide good conditions for the simulation analysis of the standard maintenance operation processes.

MTM-1 Time Prediction Method

For the simulation hours, this article uses data analyzed using JACK's Predetermined Time System (PTS). It is based on the Time Measurement Method-1 (MTM-1) theory. PTS is an internationally recognized advanced technology for time standards. Its most significant characteristic is that it is used for all kinds of actions with varying operating standards to determine the time required to complete tasks rather than only through observation or measurement. It can accurately describe the action and add the predetermined time, avoiding randomness and uncertainty that come from tests or statistical sampling. The data thus obtained are more consistent and objective than the data obtained by other methods [4, 15].

The MTM theory establishes a standard time based on carrying out repeated "basic actions". It is the most advanced and practical time measurement technology in the field of international industrial engineering. It can not only obtain accurate and objective time standards but also establish and improve working methods. This method is especially suitable for short period and high repeatability operations. The PTS system in the JACK software is mainly used as the base system of MTM theory, and it is called MTM-1. It analyzes each of the tasks of the overall process, identifies the tasks of the workers, uses the corresponding time in the basic action classification table, and then works out the time required to complete the entire process.

In civil aviation maintenance, by breaking down a set of maintenance tasks into multiple steps, we can analyze the variety and quantity of maintenance tasks. Then, by using the basic task time, the necessary time to complete the whole maintenance process can be calculated. In this way, predicting the time required to complete all the maintenance tasks during a maintenance operation and simulating the working time (T_s) needed to complete a maintenance task become possible.

Acquisition of the Simulation Work Time (T_s)

Figure 2 shows the PTS operation interface in JACK 8.4 software.

The system automatically calculates the T_s through an analysis of the dynamic simulation process and the MTM-1 theory. Then, combined with the available time which the maintenance management gave (T_a), the maintenance task-load index (TLI) is calculated as $\frac{T_s}{T_a}$.

$$\text{TLI} = \frac{T_s}{T_a}$$

Predetermined Time Standards

Task Entry | Reports | Analysis Summary

Element Number: 10 Task Number: 100 Units: metric

Task Description:

Fundamental Motion: R. Reach

Reach Task Definition

Reach Case

A. Reach to object in fixed location, or to object in other hand or on which other hand rests.

Distance Moved (cm): Hand Motion Type: 1. No hand motion at beginning or end.

Left Hand Right Hand Outside Area of Normal Vision

Add Task Update Task

Task List

| Task | Description | Code | Subtask Time - sec | Element Time - sec |
|------|-------------|------|--------------------|--------------------|
| | | | | |

Save Task File... Open Task File... Renumber Tasks Delete Selected Task

Cycle Summary

Skilled Worker Novice Worker Time Units: Seconds TMU

Total Time: 0.0 sec

Usage Dismiss

Fig. 2. The PTS time prediction tool interface in JACK

3 Case Study: Application of JACK in Removal of the Gear Wheel (for Airbus A320)

3.1 Static Environment and Dynamic Simulation Creation

JACK's static simulation creation system is also a simulation of the working environment. The creation of JACK's simulation environment includes the creation of 3D models, digital human body models and models for the relative position of each maintenance object.

For maintenance work, the entities related to the work scene mainly include various parts of the aircraft body, aircraft maintenance tools and maintenance aids. These aircraft accessories and specialized maintenance tools need to be created using an external 3D model, which is then imported into the scene using the entity import function of the JACK software. To ensure the accuracy and objectivity of the simulation, after the entities are imported into the JACK software system, it is sometimes necessary to modify and adjust the size, relative position relation, related mobile relation, and coordinate properties of all the actual datasets. Thus, a virtual work scene is set up to meet these requirements. In this paper, a preliminary 3D model of the related entities was created

with the help of the Rhinoceros 5.0 system. The parts of the 3D model are shown in Fig. 3 below.

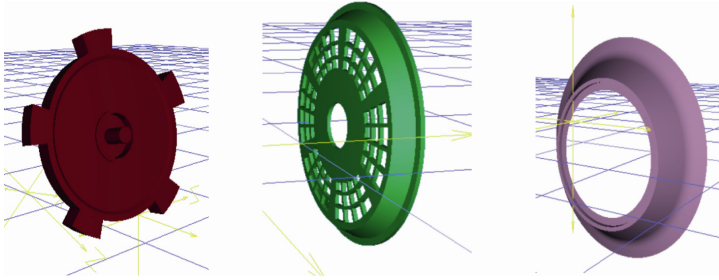


Fig. 3. A preliminary 3D model. a. Cooling fan for brake, b. cover for wheel fan, and c. annular protective cover

This paper adopts the functions of standard digital human activities in JACK and builds a model of a human civil aviation maintenance personnel's body based on Chinese anthropometric data (GB 10000-88) [3]. The digital human model is built by selecting different height, weight, and gender ratios.

All the equipment and body structures involved in the static scene are not required to be of uniform precision. To solve the difficulties associated with simulating the maintenance, we can only simulate the key attributes of the key object. Therefore, it is necessary to distinguish between the secondary features of the product model and simplify the process when the environment of the maintenance work is modeled. For example, some maintenance work associated with hangars and non-aircraft structures can be simplified, and the specific parts involved in maintenance projects can be detailed and completely simulated. Figure 4 shows a simplified version of the main wheel removal task for an Airbus A320, which was set up through the JACK static simulation module.

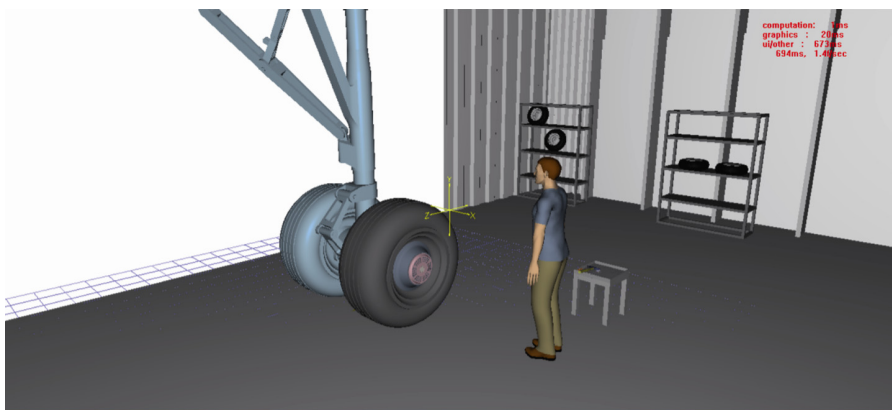


Fig. 4. Airbus A320 wheel remove operation scene

This paper gathered the maintenance actions according to the theory of kinetic element analysis [17]. According to the standard operation process document and video of the Airbus A320 wheel removal, we performed level decomposition and analyzed the four parts of the standard demolition operation process. The four parts included removing the fan cover, removing the brake cooling fan, removing the shaft nut, and removing the main wheel. This gives the A320 main tire removal disassembly task shown in Table 1 below.

Table 1. Partial task action breakdown table of the main wheel of the A320.

| Maintenance events | Maintenance work | | Breakdown of maintenance work | The basic dynamic element |
|----------------------------|------------------|--------------------------------|-----------------------------------|---|
| | Sort | Name | | |
| Main wheel removal of A320 | 1 | Removing the fan cover | Loosen the screws | Take (tool) → Position → |
| | | | Remove the screws | Pressure → Position → |
| | | | Remove the fan cover | Put down → Take (fan cover) → Put down |
| | 2 | Removing the brake cooling fan | Remove the fuse | Take (tool) → Position → |
| | | | Loosen the screw nut | Pressure → Position → |
| | | | Remove the brake cooling fan | Put down → Take (Brake cooling fan) → Put down |
| | 3 | Removing the shaft nut | Take off the open sale insurance | Take (tool) → Position → |
| | | | Loosen the shaft nut | Pressure → Put down → |
| | | | Remove the shaft nut | Take (tool) → Position → Pressure → Position → Put down → Take (shaft nut) → Put down |
| | 4 | Removing the gear wheel | Install the wheel shaft protector | Take (wheel shaft protector) → Position → |
| | | | Remove the gear wheel | Pressure → Position → Posture (stand up) → Take (gear wheel) → Pressure → Position → Go |

After obtaining the basic moving sequence, the TSB function in JACK8.4 was used to create a dynamic simulation of those four processes. Figure 5 shows screenshots based on the dynamic simulation video.

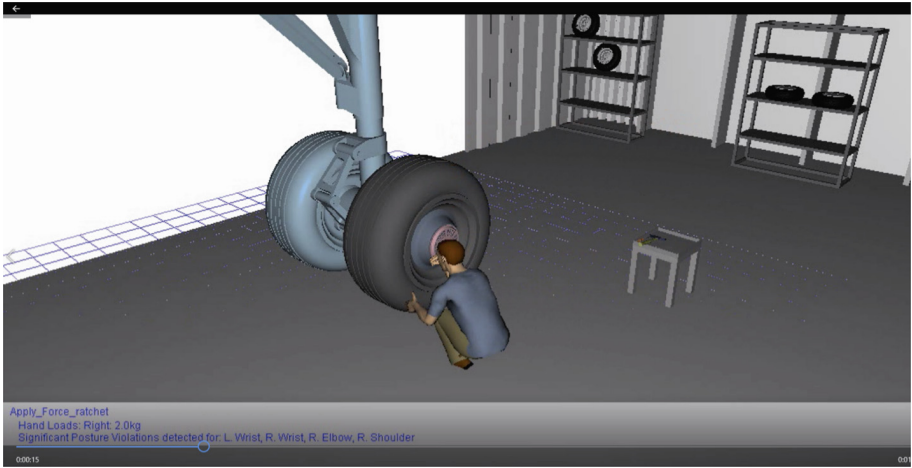


Fig. 5. Dynamic simulation video screenshot

3.2 Task-Load Evaluation Parameter Acquisition

JACK can calculate and analyze the following data according to the relevant motion data of the dynamic simulation:

Figure 6 shows that the simulation working time T_s is 72 s.

| -----Analysis Details----- | |
|----------------------------|-----------|
| Figure: | human |
| Weight: | 77.690 kg |
| Height: | 175.49 cm |
| Gender: | Male |
| Cycle Time (sec): | 72.064 |

Fig. 6. TSB simulation time analysis of CSV format output

According to the investigation, the time T_a of maintenance management for the above maintenance tasks is 3 min, i.e., 180 s.

3.3 Calculation and Analysis of Task-Load Evaluation Results

According to the parameters of the TSB and TAT analysis tools, the task-load of the gear wheel removal task in the Airbus A320 was obtained. Using the task-load evaluation model, the detailed calculation process was as follows:

$$TLI = \frac{T_s}{T_a} = \frac{72}{180} = 0.4 \quad (2)$$

The calculation result shows that the task-load of the gear wheel of the Airbus A320 is acceptable and can even be further optimized in terms of time utilization. For example, when making the maintenance work plan, the planned working time can be shortened and changed from 3 min to 2.5 min. This effectively improves time utilization while ensuring that the task-load is within acceptable limits.

4 Discussion and Conclusion

This paper combines the practical work of civil aviation personnel with time utilization as a basic idea to measure task-load. To evaluate the technical implementation, the paper creatively establishes a dynamic simulation method for actual maintenance work using TSB in JACK. Thus, by obtaining a task-load evaluation for each parameter, this method effectively solves the disadvantages from previous task-load evaluation methods, such as strong subjectivity, poor portability, and a requirement of strict laboratory conditions. In the evaluation of maintenance workloads, this paper applied a method of task-load measurement based on a time occupancy rate similar to DORATASK.

However, in practical applications, other factors that may affect the actual work should also be taken into account, such as the energy consumption rate, strength load, attitude load, and various natural environmental factors. Future research on the maintenance workload evaluation should focus on these factors.

References

1. CAAC: Guidelines for the Human Resource Assessment of Airborne Carrier Pilots, 6 August 2014
2. CAAC: Maintenance Time Management, 7 November 2011
3. Ding, Y.L., Cheng, G.P.: Ergonomics. Beijing Institute of Technology Press, Beijing (2013). (in Chinese)
4. Edward, J.: Methods Analysis and Work Measurement. McGraw-Hill, New York (1984)
5. Guo, D.: Aviation Maintenance Ergonomics, pp. 116–135. National Defend Industry Press, Beijing (2007). (in Chinese)
6. Guo, F., Qian, S.S.: Human Factors Engineering, pp. 148–153. China Machine Press, Beijing (2005). (in Chinese)
7. Han, S.C., Hu, M.H., Jiang, B.: The study of relationship between sector capacity and ATC controller's workload. *Air Traffic Manag.* **6**, 42–45 (2000). (in Chinese)
8. Hua, Y.C., Sun, C.L.: Human Factors in Aviation Maintenance and Their Application, pp. 51–77. China Civil Aviation Press, Beijing (2009). (in Chinese)
9. International Civil Aviation Organization: Air Traffic Service Planning Manual (DOC9426). International Civil Aviation Organization, Montreal (1984)
10. Karger, D.W., Bayha, F.H.: Engineered Work Measurement: the Principles, Techniques, and Data of Methods-Time Measurement, Background and Foundations of Work Measurement and Methods-Time Measurement, Plus Other Related Material. Industrial Press, New York (1987)
11. Liu, X.Y., Li, Y.J., Jiang, L.: Study on practitioners' workload in civil aviation. *China Saf. Sci. J.* **18**(6), 28–33 (2008)

12. Ren, Q.J.: License planning for airline cabin-crews and production planning for dispatchers. A dissertation for doctor's degree, University of Science and Technology of China (2013)
13. Sun, R.S., Hu, Z., Wang, L., Huangfu, G.X.: Influencing factors of maintenance personnel fatigue based on WLSN and entropy weight method. *Saf. Environ. Eng.* **23**(3), 167–170 (2016). (in Chinese)
14. Tofukuji, N.: An airspace design and evaluation of enroute a sector by air traffic control simulation experiments. *Trans. Inst. Electron. Inf. Commun. Eng.* **78**, 358–365 (1995)
15. William, H.C.: In-flight and post flight assessment of pilot workload in commercial transport aircraft using the subjective workload assessment technique. *Int. J. Aviat. Psychol.* **2**(2), 77–93 (1992)
16. Yuan, L.P., Sun, R.S., Liu, L.: Measuring the workload of the air traffic controller based on DORATASK method. *J. Saf. Environ.* **14**(3), 76–79 (2014). (in Chinese)
17. Zhang, Y.H.: Research on Operations Simulation and Evaluation Method of Ergonomics for Virtual Assembly. Harbin Institute of Technology (2015). (in Chinese)
18. Zhao, Y.: Research on the evaluation and prediction of the human's workload capacities for the in-orbit manipulation. National University of Defense Technology (2014). (in Chinese)



The Identification of Human Errors in the Power Dispatching Based on the TRACEr Method

Xiaobi Teng², Yanyu Lu¹(✉), Zhen Wang¹, Bingbing Song², Hai Ye²,
Yi Zhou², and Shan Fu¹

¹ School of Electronics, Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
luyanyu@sjtu.edu.cn

² East China Branch of State Grid Corporation of China,
Shanghai 200120, China

Abstract. Most of the power dispatching accidents were caused by human errors. Human error should be symptoms of systemic problems and opportunities to learn about the features of complex systems. Therefore, the identification and analysis of the human errors in the power dispatching is the significant to guide against the human risk and ensure the stable and safe operation of power nets. Human error identification methods have been used to identify the nature of the human errors and causal factors, and recovery strategies in many industrial domains such as the aviation, nuclear power and chemical processing industries. The Technique for Retrospective and Predictive Analysis of Cognitive Errors (TRACEr) is a human error identification technique that was developed for use in the air traffic control domain. In this study, the TRACEr was improved in the combination of the task features of the power dispatching and human information processing, and was used to identify the human errors in the power dispatching. A total of seventy-two incidents or accidents performed by operators were analyzed. The analyzing processing was carried out with the objective of classifying task error, identifying external error modes, internal error modes and psychological error mechanisms, and identifying the performance shaping factors. The performance factors analysis considered the time, interface, training and experience, procedures, organization, stress and complexity which may have an impact to the task and help to propose some recovery strategies. The results revealed that the identification was a necessary and effective step toward the safety improvement of power dispatching.

Keywords: Human factors · Power dispatching · Human error identification

1 Introduction

The safety and reliability of the power system operation are critical issues for maintaining stable electricity supply, ensuring economic growth and guaranteeing people's normal life order. With the development of the technologies in the power system, increasingly sophisticated automation has been introduced into the power system

operation, including power dispatching. To a large extent, safety benefit from the increasingly the reliable automation. However, the system complexity and lack of the transparency put forward higher requirement on the dispatchers. Statistically, about 75% of the accidents in the power dispatching operations were attribute to the human factors [1]. Therefore, the identification and analysis of the human errors in the power dispatching is the significant to guide against the human risk and ensure the stable and safe operation of power nets.

In addition to the power dispatching, human error has been considered as a significant factor in the incidents and accidents in complex systems, such as nuclear power and civil aviation [2]. Reason defined human error as “All those occasions in which a planned sequence of mental or physical activities fails to achieve its intended outcome, and when these failures cannot be attributed to the intervention of some chance agency”. Furthermore, Reason proposed the classification of human error, including the slips, mistakes and violations [3].

The human error identification methods are widely investigated in the complex systems. In the new view, researchers think that human error is not a cause of an incident or accident. It is the consequence, the effect, the symptom of the accident deeper in the whole system [4]. Human error provides information to help diagnosing the systems. Therefore, human error identification methods are developed to identify the nature of operator errors and causal factors, recovery strategies.

Several human error identification methods have been developed in the different domains. The systematic human error reduction and prediction approach (SHERPA) was developed for the nuclear reprocessing industry, which is a classification method to identify potential errors associated with human activity [2]. The human error template (HET) method was developed for the civil flight deck [5]. The hazard and operability study (HAZOP) method was first developed by ICI for the safety of a plant or operation [6, 7]. The Cognitive Reliability and Error Analysis Method (CREAM) was developed for an analysis of the human reliability analysis approaches, which can be used both to predict potential human error and to analyze error [8]. The Human Error Identification in Systems Tool (HEIST) adopted a series of error identifier prompts to identify potential errors [9]. The Human Factors Analysis and Classification System (HFACS) was developed to investigate and analyze human error in aviation based on the “Swiss cheese” model of accident causation [10, 11].

The technique for the retrospective analysis of cognitive errors (TRACER) was developed specifically for human error identification in the air traffic control (ATC) domain, which can be used either proactively to predict potential error and analyze operators’ performance or retrospectively to investigate accidents [12]. The method combines the psychological, physical and external factors based on the experiment and applied psychology, human factors and communication theory. Moreover, the TRACER method has been applied in the railway domain [13], ship accident [14] and maritime transportation industry [15].

In this paper, the TRACER method was used to identify and analyze a set of dispatching accidents in the power system in consideration of similarity of the tasks in power dispatching and ATC. The objective is to characterize dispatching incidents and accidents in terms of task errors, human-machine interface and cognitive domains involved the accidents.

2 TRACER Method

TRACER method is focused on the human-machine interface and the cognitive processes of the operator. According to the TRACER, some environmental or situational factors influence the operator's mental state, which causes the failure of the cognitive processes, and finally lead to an accident. Therefore, it does not only analyze the external and observable manifestation of the task error but goes deep in the cognitive domain that help analyst to explore the context that lead the operator make errors.

According to Shorrock and Kirwan [12], TRACER method has a modular structure with various layers: Task Error, External Error Modes (EEMs), Internal Error Modes (IEMs), Psychological Error Mechanisms (PEMs), Performance Shaping Factors (PSFs).

The TRACER method was used in this study to identify the human error in the power dispatching as follows:

1. Defining the task error, such as communication error, material check error, monitoring error.
2. Defining the error or violation.
3. Identifying the external error modes. Table 1 presents the EEM taxonomy.

Table 1. External error mode taxonomy.

| Timing and sequence | Selection and quality | Information transfer |
|---------------------|------------------------------|-----------------------------|
| Action too early | Omission | Unclear info transmitted |
| Action too late | Action too much | Unclear info recorded |
| Action too long | Action too little | Info not obtained |
| Action too short | Action in wrong direction | Info not transmitted |
| Action repeated | Wrong action on right object | Info not recorded |
| Mis-ordering | Right action on wrong object | Incomplete info transmitted |
| | Wrong action on wrong object | Incomplete info recorded |
| | Extraneous act | Incorrect info transmitted |
| | | Incorrect info recorded |

4. Identifying the failure of cognitive domains. The four cognitive domains comprise perception, memory, planning and decision-making and action execution.
5. Identifying internal error modes and psychological error mechanisms. IEMs describe what cognitive function failed or could fail, and in what way, and provide an interface between EEMs, PEMs, and the cognitive domains, and thus give an intermediate level of detail. For example, the 'perception' was divided into 'visual' and 'auditory'. PEMs describe the psychological nature of the IEMs, such as 'expectation bias', 'perceptual confusion' and 'distraction' in 'perception' domain.
6. Identifying the performance shaping factors. In the study, PSFs included time, interface, training and experience, procedures, organization, stress, and complexity. PSF categories and associated keywords are presented in Table 2.

Table 2. Psychological error mechanisms taxonomy.

| Category | Examples |
|--------------|--|
| Time | Emergency tasks; night shift |
| Interface | No information; unclear information; conflicting information |
| Training | No enough training or experience |
| Procedures | No procedure; fuzzy procedure; too simple procedure; wrong procedure; unreadable procedure |
| Organization | Insufficient personnel; insufficient cooperation; poor working environment |
| Stress | High workload/stress; fatigue |
| Complexity | New task; complex task |

3 Accident Analysis

In this study, 72 incidents and accidents reports have been analyzed using the TRACER method. The accident reports came from State Grid East China Electric Power Control Center and covered dispatching accidents in a period from 2015 to 2017. Since there was more than one error in an incident or accident, the analysis have produced 113 task errors using the TRACER method described in the previous section. The flowchart of human error analysis was presented as Fig. 1.

3.1 External Error Modes

The TRACER method provides 3 categories for the external error modes as mentioned in the previous section (timing and sequence, selection and quality, and information transfer). The main EEM is information transfer with a percentage of 46.05%, while the percentage of the ‘timing and sequence’ and ‘selection and quality’ were 35.53% and 18.42%, respectively, as shown in Fig. 2.

The main error mode in the category ‘information transfer’ was ‘incorrect information transmitted’ (13.16%, Fig. 3). For example, the operator transmitted incorrect electrical generation or peak values. The second error mode was ‘incomplete information transmitted’ (10.53%), such as transmitting incomplete repair schedules. The category ‘information not transmitted’ was also an important error mode (9.21%). The operator in the provincial power grid or power station forgot to transmit the failures of the system or transmit their maintenance activities. In addition, ‘incorrect info recorded’ occupied the percentage of 6.58%, such as recording wrong device. Other relevant EEMs were ‘unclear info transmitted’ (2.63%), ‘info not obtained’ (2.63%) and ‘incomplete info recorded’ (1.32%).

The ‘action too early’ was the most among all the EEMs with a percentage of 19.74%, as shown in Fig. 4. when a failure happened, operators often took actions other than reporting to the Control Center as required. The ‘action too late’ (7.89%) typically involved detecting the warning too late, and the ‘action too long’ (6.58%) involved that operator did not process failures within the required time.

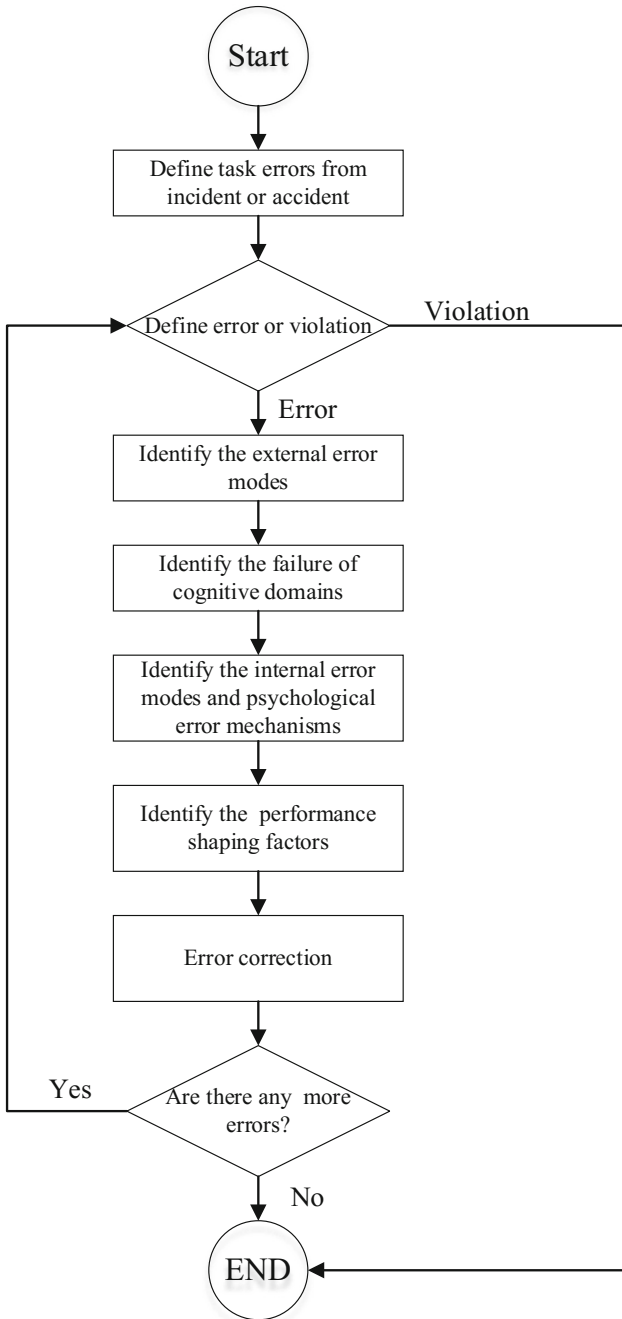


Fig. 1. Flowchart of the error analysis in the study

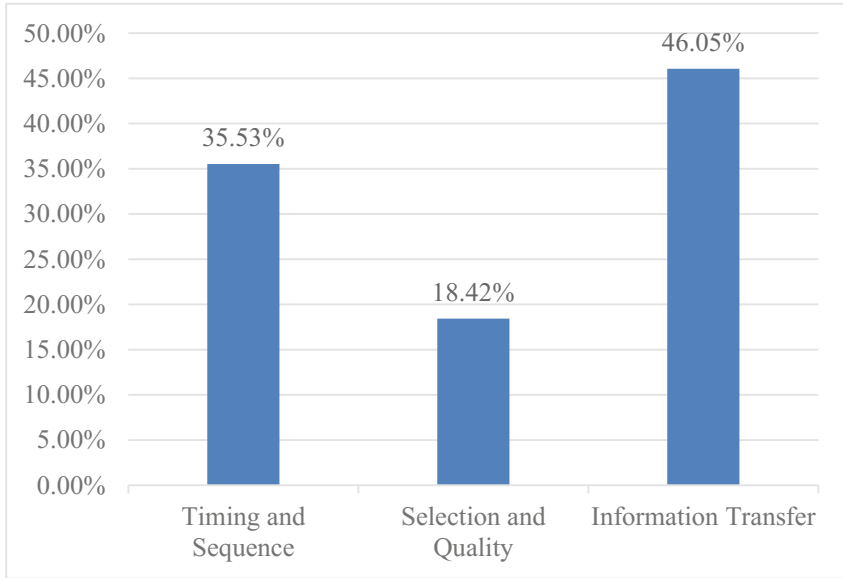


Fig. 2. External error modes

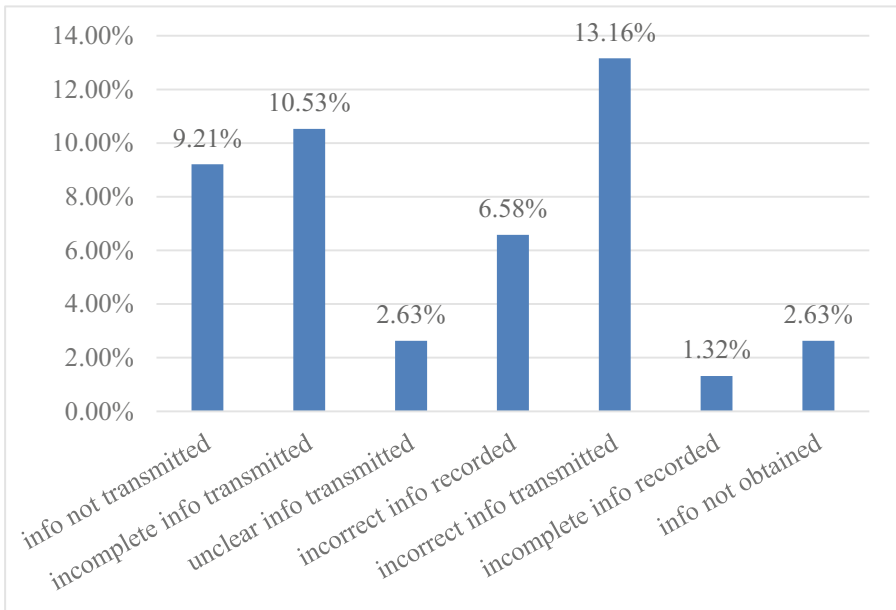


Fig. 3. Percentage of error modes in 'information transfer' category

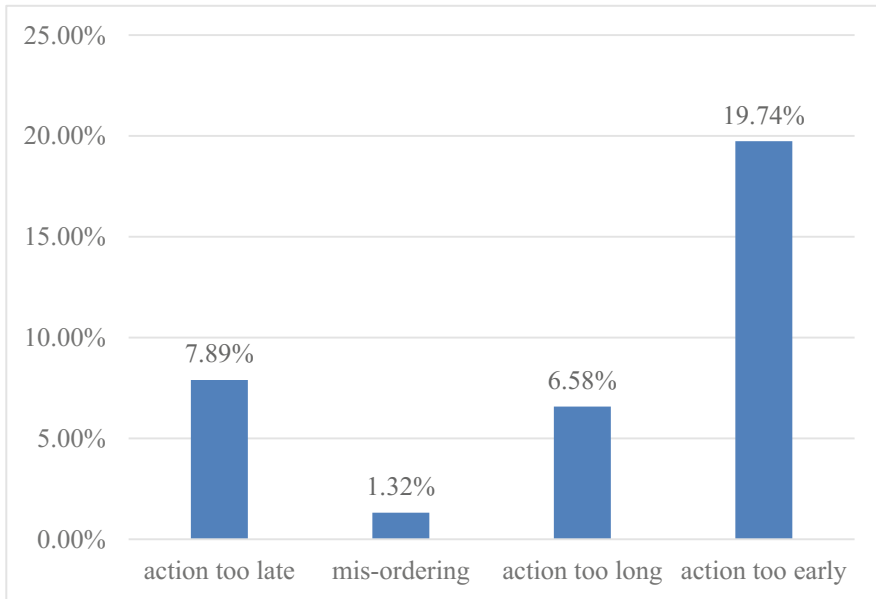


Fig. 4. Percentage of error modes in ‘timing and sequence’ category

The ‘selection and quality’ category had main error mode of ‘omission’ with a percentage of 9.21% (Fig. 5). Operators often omitted some information when formatting the operation tickets. Other relevant error modes were action in wrong direction (3.95%), action too little (2.63%), right action on wrong object (1.32%) and extraneous act (1.32%).

3.2 Cognitive Domains

The analysis in the study showed that the cognitive domains related to the task errors were 35.40% of planning and decision-making, 24.78% of action execution, 17.70% of Memory, 11.50% of Perception, and 10.62% of Violation, as shown in Fig. 6. Obviously, the ‘planning and decision-making’ domain was the most failure in the cognitive processing, meaning the operator had a worse situation awareness.

3.3 Performance Shaping Factors

Figure 7 showed the performance shaping factors that might influence the operators’ performance and result in errors. The insufficient information in the interface and stress were the main factors with the percentage of 17.05%. The second factor was night shift (16.28%). Moreover, as regarding to the organization factor, the contributors were insufficient personal (10.85%), insufficient cooperation (4.65%) and poor environment (3.10%), respectively.

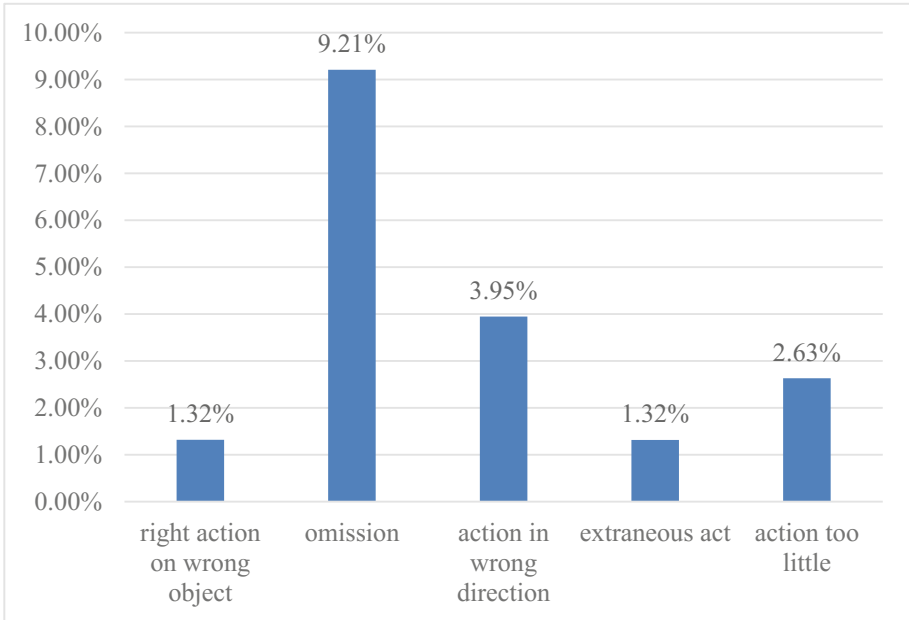


Fig. 5. Percentage of error modes in 'selection and quality' category

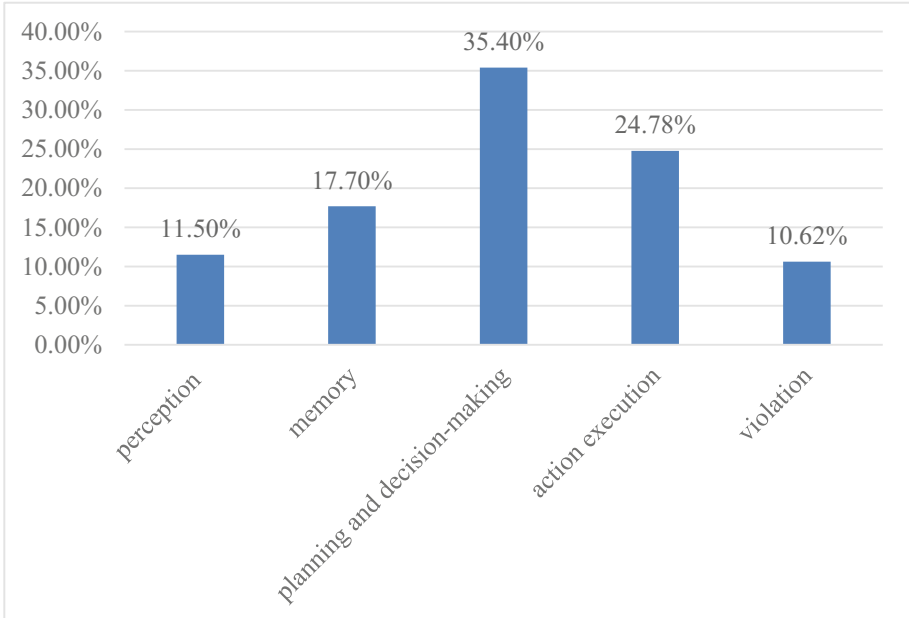


Fig. 6. Percentages of cognitive domains in related to task errors

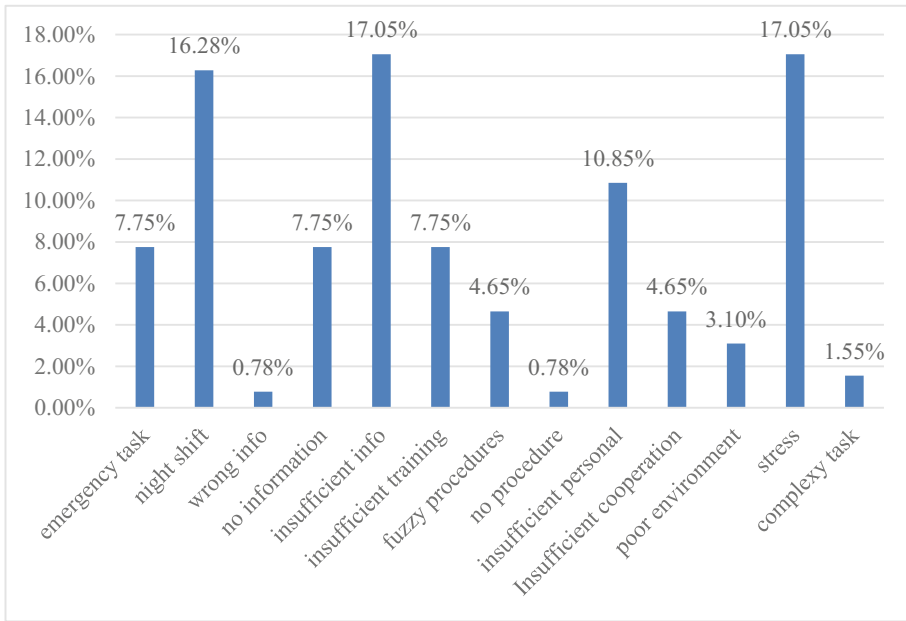


Fig. 7. Main performance shaping factors that might cause the task errors

4 Discussions and Conclusions

The objective of this study was to demonstrate the utility of the TRACER method for analyze the operators' error in the power dispatching. The results revealed that the method can benefit the identification of the operators' cognitive state and the exploration of the context that influences the operators' performance through EEM, IEM, PEM and PSF analysis.

The TRACER method integrating Wickens model of information processing [16] into its model is a structured analysis procedure which is useful for the analyst to track and classify errors through the stages of human information processing. In the study, the 'planning and decision-making' was identified as the main cognitive domain related to the task errors. It is attributed to the task characteristic of the power dispatching that an operators in some place seldom get complete information about the state of the power grid, and available, since he typically deals with only a few functional areas of power system operations [17]. Therefore, enhancing the operators' situation awareness is a critical issue for the power dispatching.

The TRACER method considered the PSFs within the whole system which contributed to errors identified and error correction. In the study, 'night shift', 'insufficient information in the interface' and 'stress' were considered as main factors that might have impact on the operators' performance. The results suggested that the human-machine interface, task assignment and other relevant factor should be adapted.

As mentioned above, the TRACER method which was developed based on psychological theory could require high training time and an understanding of psychology

in order to use the method. Meanwhile, the method highly relies on the background, experience, and knowledge of the domain and task analyzed.

To conclude, the paper demonstrated that the TRACER method can be applied as a retrospective analyzing tools for the incidents and accidents in the power dispatching. The results might benefit the proposal of some recovery strategies and improvement of the operation safety for the power dispatching. However, the validity and reliability of the method still are required to explore in combined with detecting operators' cognitive state using objective measurements.

References

1. Yang, F., Wu, C., Wang, F., Ma, S.: Review of studies on human reliability researches during 1998 to 2008. *Sci. Technol. Rev.* **27**(8), 87–94 (2009)
2. Stanton, N.A., Salmon, P.M., Rafferty, L.A., Walker, G.H., Baber, C.: *Human Factors Methods: A Practical Guide for Engineering and Design*. Ashgate Publishing Limited, Farnham (2013)
3. Reason, J.: *Human Error*. Cambridge University Press, Cambridge (1990)
4. Dekker, S.: *The Field Guide to Understanding Human Error*. Ashgate Publishing Company, Farnham (2006)
5. Marshall, A., Stanton, N., Young, M., Salmon, P., Harris, D., Demagalski, J., Waldmann, T., Dekker, S.: Development of the human error template—a new methodology for assessing design induced errors on aircraft flight decks
6. Kletz, T.A.: HAZOP and HAZAN: notes on the identification and assessment of hazards. *J. Hazard. Mater.* **8**(4), 385–386 (1984)
7. Swann, C.D., Preston, M.L.: Twenty-five years of HAZOPs. *J. Loss Prev. Process Ind.* **8**(6), 349–353 (1995)
8. Hollnagel, E.: *Cognitive Reliability and Error Analysis Method (CREAM)*. Elsevier, Oxford (1998)
9. Kirwan, B.: A guide to practical human reliability assessment. *Int. J. Ind. Ergon.* **17**(1), 69 (1994)
10. Shappell, S.A., Wiegmann, D.A.: A human error approach to accident investigation: the taxonomy of unsafe operations. *Int. J. Aviat. Psychol.* **7**(4), 269–291 (1997)
11. Shappell, S.A., Wiegmann, D.A.: The human factors analysis and classification system-HFACS. *Am. Libr.* **1**(1), 20–46 (2000)
12. 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)
13. 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)
14. Hofmann, S., Schröder-Hinrichs, J.U.: *CyClaDes Task 1.2 Incident and accident analysis*. Document ID Code: CY112. 00.02. 041.041, WMU (2013)
15. Graziano, A., Teixeira, A.P., Soares, C.G.: Classification of human errors in grounding and collision accidents using the TRACER taxonomy. *Saf. Sci.* **86**, 245–257 (2016)
16. Wickens, C.D.: *Engineering Psychology and Human Performance*. Charles E. Merrill Publishing Company, Columbus (1992)
17. Guttromson, R.T., Schur, A., Greitzer, F.L., Paget, M.L.: Human factors for situation assessment in grid operations (2007)



Ergonomic Evaluation Study of Occupant Function Allocation for Riot Vehicle Based on Task Load

Qun Wang^{1(✉)}, Fang Xie¹, Runing Lin¹, Xiaoping Jin², and Xue Shi¹

¹ General Technology Department, China North Vehicle Research Institute, Beijing 100072, China
buaawq1988@163.com

² College of Engineering, China Agricultural University, Beijing 100083, China

Abstract. With the development of intelligent equipment, the police riot vehicle (PRV) cabin space become smaller, but the mission demand does not reduce, but increase. At present, there is no mature unmanned technology, so the cabin designer has to reduce the number of occupants and redistribute occupant functions. In order to test the feasibility of the new system, the new scheme of occupant function allocation must be evaluated at first. This study proposes a multi-index evaluation method to the new PRV occupant function allocation solution based on task load. A semi-physical engineering simulator was built and ergonomic experiments that combined subjective and objective evaluation to the new occupant function allocation scheme were performed in the simulator. The experiment task was two typical mission profiles: one is the long-range target attack; another is the close-range target attack. The task load includes subjective and objective evaluation indexes. The evaluation questionnaire was designed based on the multi-resource theory that contained 5 channels: visual, auditory, cognitive, speech and movement and the workload of each channel was given with a subjective scale. The objective indexes consist of the main task performance, the sub-task performance of detection response task (DRT) and two physiological indexes of skin electricity and heart rate variability. The experimental results showed that the change of occupant numbers has an impact on the performance of different task stage. The result of the subjective evaluation showed that the total task workload score were highly related to the workload score of each channel of visual, cognitive and movement channel. The conclusion of this study can be made that the change of the occupant number has a greater impact on the occupant performance with an increased physical load within the acceptable range; The evaluation method based on task load is feasible in the semi-physical simulator environment. The study can provide theoretical support for the evaluation of the occupant assignment function in the new type of PRV.

Keywords: Ergonomic evaluation · Function allocation · Mental workload

1 Introduction

The change in the operation function of the occupant in the police riot vehicle (PRV) is the result of exploitation and evaluation of human's ability in task. Occupant's function

has great changes compared with the initial design of PRV. Occupant's function varies from operation to information processing and decision-making. In addition, coordination between vehicles becomes more and more important in the future, which may lead to decrease of occupant's number and tightly integration of information interface. However, people's ability especially information processing and analyzing has a certain limit [1]. Vehicle occupants also need to pay attention to changes in vehicle information during driving, and deal with a large number of unpredictable complex information caused by external environment quickly. Overload situation appears once occupant is assigned tasks which beyond his or her ability. Overload situation leads to errors, thus affecting the safety of occupants and vehicles.

Therefore, occupant's function should be assigned according to occupant's ability and workload impact factors, and determine optimal function allocation based on workload evaluation results. Occupant is freed from the complicated routine operation, and individual ability focus on information monitoring, collection, analysis and decision making. So, vehicle can play the best role to ensure the completion of the task performance.

Most research on occupant workload focus on mental workload. Mental workload is a multidimensional concept, which is closely related to the information processing. Mental workload involves the work requirement, time pressure, task complexity, operator's ability, operator's effort and so on [2]. The main measurement methods are director measurement method, sub-task measurement method, subjective evaluation method and physiological measurement method [3].

According to the definition of cognitive load, task load is closely related to information received when the task is executed. Therefore, this paper proposes an objective evaluation method of task load for operational task flow based on multiple resources theory [4], which combines with task performance measurement methods, subjective evaluation methods and physiological measurement techniques. The method studies the change of task load before and after the change of occupant's function based on vehicle semi-physical simulator, which provides reference for the function allocation of occupant.

2 Method

2.1 Task Design

Long range target attack and close target attack are selected to verify availability of method.

Except for distance to attack target, two types of tasks have difference in operator performed attack operation in the 3-occupant mode. In the long range target attack mode, Gunner is responsible for receiving instructions from commander, aiming at long distance target and attacking target. In the close target mode, commander is responsible for attacking close target [5].

Except for distance to attack target, two types of tasks have difference in attack operation, the attack operation performed by gunner is originally assigned to commander in the 2-occupant mode. The task load of commander is considered much larger than it

in 3-occupant mode before experiment. Semi-physical simulator experiment need to be conducted to evaluate task load and determine whether high task load influencing task operation. Commander is responsible for assigning instructions, aiming at long range target, attacking target and reporting to superior in the long range target attack mode. Commander is responsible for attacking close target in the close target attack mode.

2.2 Experiment Design

2-occupant experiment adopts experiment design with two independent variables.

Occupant group. between-subjects designed, there are two levels including driver and commander.

Task type. within-subjects designed, there are two levels. Every subject in groups performs long range attack task and close attack task, each level repeat twice in turn. The order of task is long range attack task, close attack task, close attack task and long range attack task. Experiment design shows in Table 1.

Table 1. Experiment design

| Independent variable: task type | Independent variable: occupant | |
|-----------------------------------|--------------------------------|--------------------|
| | Level 1: driver | Level 2: commander |
| Level 1: long range target attack | Condition 1 | Condition 2 |
| Level 2: close target attack | Condition 3 | Condition 4 |
| Remark | Without DRT | With DRT |

2.3 Subjects

Subjects are healthy PRV drivers. They have no eye disease and their vision or corrected visual acuity in 1.2 or more. Subjects are in good health condition during the experiment. There are 30 male subjects who are randomly divided into commander and drivers groups averagely.

2.4 Experiment Apparatus

The experiment apparatus include a vehicle semi-physical simulator, a camera, a physiological feedback instrument and a detection response task (DRT) [5]. Table 2 shows their functions.

Table 2. Experiment apparatus and functions

| Name of apparatus | Function of apparatus |
|-----------------------------------|--|
| Vehicle semi-physical simulator | Simulate the operation of the vehicle man-machine interface |
| Camera | Record the whole operation processes of occupants, calculate occupants' performance after experiment (accuracy and response time of target attack) |
| Physiological feedback instrument | Record the galvanic skin response and the heart rate variability of occupants |
| Detection response task (DRT) | Detect response time of sub-task and miss rate |

The subjective evaluation is carried out after each experiment. The subjects rate overall load and 5-channel load respectively from 0 to 10 in different task type and different occupant type. The 5-channel indicates visual channel, auditory channel, cognitive channel, behavioral channel and oral channel. Higher score represents the heavier load.

2.5 Experiment Process

In order to master the use of experiment apparatus, subjects are trained to be familiar with the operation of DRT equipment and the vehicle semi-physical simulator before the start of the experiment. Then, subjects are asked to perform a number of tasks related to the experiment, and let subjects wear apparatus of physiological and DRT after finishing above operation. The experiment is carried out according to the above experimental design. Report "Attention, task complete, please assign next task" to the lab after completing each task type, and subjects process to the next task experiment when they hear "task continue", and it is the time to finish the experiment until subjects hear the voice of "End experiment". After completing each experiment, subjects are asked to mark for subjective rating scale.

3 Results and Discussion

3.1 Data Description

Extract index for the single task type experiment. The load-research index includes two type of physiological index and subjective evaluating data for 5-channel load, one type index include occupant's primary-task performance (second, s), sub-task (cognitive attention) performance (response time, ms; miss rate, %), the galvanic skin response, and another type is the heart rate variability. Occupant act coherently in experiment, therefore only test commander's load. Various task-moments during the target attack are depicted as the Fig. 1.



Fig. 1. Each task phase during target attack

The time points which are required to record are commander spotting target, commander accepting attack-instruction from superior attacking target successful (when hearing cannon’s sound) and commander restoring driving instruction (including reporting to superior and ordering occupant). Gather commander’s accurately data of attacking target from using camera after the experiment. When it comes to the lack of data results from experiment apparatus and metering mistake, we adopt mean-value invariant principle to substitute the same tested metering data in corresponding metering group. The extra data need to gather show in the below Table 3.

Table 3. Commander’s task load evaluation data

| Data type | Evaluating index |
|-------------------------------|---|
| Long range target attack task | Completion time of the task phase 1 |
| | Completion time and accuracy rate of the task phase 2 |
| | Completion time of the task phase 3 |
| Close target attack task | Completion time of the task phase 1 |
| | Completion time and accuracy rate of the task phase 2 |
| | Completion time of the task phase 3 |
| Cognition attention data | Response time |
| | Miss rate |
| Physiological signal | Heart rate |
| | Galvanic skin |
| Subjective evaluation | Commander’s load score |

3.2 Data Analysis Results and Discussion

The data analysis focuses on the influence for the commander’s load after merging the gunner and commander’s duty. The analysis results are summarized from three aspects. The first aspect focus on the comparative analysis between the task performance of the 3 occupants or 2 occupants and DRT measurement data, and this part mainly discuss the influence of different task types and number of occupant on the task load of commander. In the second aspect, we accomplish a comprehensive analysis about the change of task load in the process of commander from physiological index and subjective score. In the third aspect, the subjective score is taken as the object of analysis, and the

relationship between the variation regulation of different channels' rate with the process of the task and the total load rating is discussed.

Task Performance and DRT Measurement Data Analysis. The completion time of each task stage in different task type and the quantifiable load value from DRT are shown in Table 4.

Table 4. Task performance and DRT load quantitative statistical results (mean)

| Task | Task performance (s) | | | | | | Reaction time (ms) | | Miss rate (%) | |
|--|----------------------|------------------|------|------|------|------|------------------------------|--------|---------------|------|
| | 3 | | | 2 | | | 3 | 2 | 3 | 2 |
| Number of occupant | | | | | | | | | | |
| Task phase | 1 | 2 | 3 | 1 | 2 | 2 | – | – | – | – |
| Long-range target attack task | 9.79 | (Gunner : 50.44) | 8.51 | 3.68 | 32.5 | 3.36 | 169.40 | 753.48 | 16.3 | 56.6 |
| Close target attack task | 6.94 | 15.5 | 7.92 | 1.47 | 13.6 | – | 665.39 | 894.09 | 63.6 | 56.0 |
| <i>Main effect (DRT Cognition attention index)</i> | | | | | | | | | | |
| Task type | Reaction time (ms) | | | | | | F(1, 109) = 8.618, p = 0.004 | | | |
| | Miss rate (%) | | | | | | F(1, 109) = 13.30, p < 0.001 | | | |
| Number of occupant | Reaction time (ms) | | | | | | F(1, 109) = 14.05, p < 0.001 | | | |
| | Miss rate (%) | | | | | | F(1, 109) = 6.05, p = 0.012 | | | |

1. Primary-Task Performance Data Analysis. From the performance result, the change in the number of occupants has an impact on the different tasks performance. In the second phase that implement the long range target attack, we can see that the commander task performance of the 2-occupant is significantly higher than the task performance of gunner in the 3-occupant ($32.5\text{ s} < 50.44\text{ s}$, $F(1, 54) = 5.5$, $p = 0.023 < 0.05$). It illustrates that the 2-occupant mode completed faster than the 3-occupant mode in the task, and the performance of commander has been improved in the task phase 1 and phase 3.

The long range target attack is more difficult and time-consuming than the close target attack at the 2-occupant mode. This is not only reflected in the completion time of the attack task, but also in the response time of the commander order task. The time spent by the commander assign task in the long range target attack mode is significantly higher than the close mode ($F(1, 53) = 9.36$, $p = 0.003 < 0.05$).

2. DRT Measurement Data Analysis. From the result of DRT measurement data (Fig. 2), the number of occupants on the influence of commander's load is associated with task types. The task load of commander is significantly increased in two different task types, and this phenomenon is more obvious in the long range target attack task type. Attacking task is performed by gunner in the 3-occupant mode, and the executor is commander if in the

2-occupant mode. This resulted in the markedly increasing of the task load, which it doesn't cause any decreasing of the task performance, it improves instead.

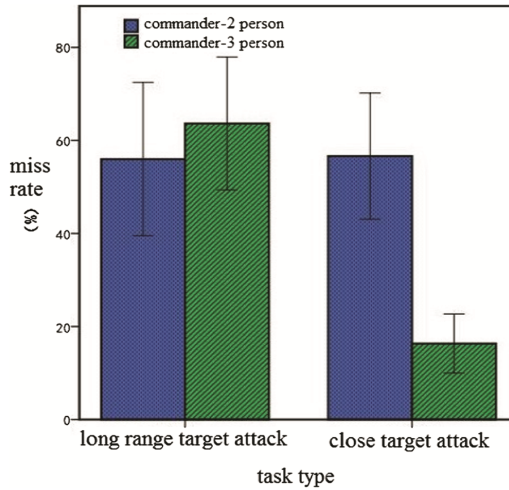


Fig. 2. Commander's DRT load analysis under different occupants (95% confidence interval)

The time of target search is under the influence of task type and the number of occupants, as Fig. 3. In close target attack task type, the decreasing of the number of commander result the observably increasing of the search time. In long range target attack task type, the decreasing of the occupant number result the decreasing of the search time. So the interaction between task type and the number of occupant on the influence of target search time is observably. Task type has dominant impact on the target search time, and the search time of long range target is higher than the search time of close target ($F(1, 176) = 9.80, p = 0.002 < 0.01$) in both two different mode of the occupant's number.

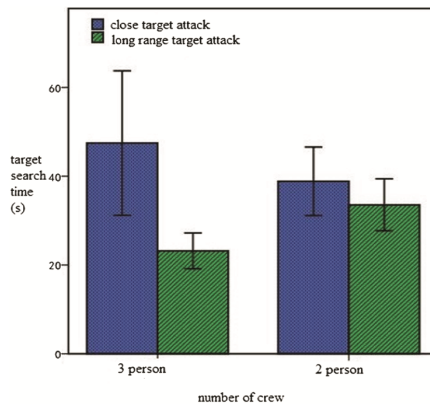


Fig. 3. The time of target search (95% confidence interval)

The Change Law of Physiological Index Data During the Process of Tasks. The physiological index extract the original heart rate and the GSR (galvanic skin response) signal. The SCL index in GSR signal, the LF/HF index in the HRV (heart rate variability) and the Pulse can effectively represent the degree of the task-performing person's load. The more value-increasing compared to the baseline suggests the larger load [6].

In order to eliminate the time effect and the individual difference, we use the normalization data to process when analyze the load of task-performing persons in the different task type, and normalization proceed in the commander or gunner's task type measured data, proceed Z-score standardization on one index of single event time order.

The normalization of commander physical indexes is shown in Fig. 4. The values are standardized indicators, if one is greater than zero, it suggests that it is greater than the average indicator of the whole task phase, if one is less than zero, it suggests that it is less than the average indicator of the whole task phase. The load variation of the whole task represents a high-to-low trend on the time order.

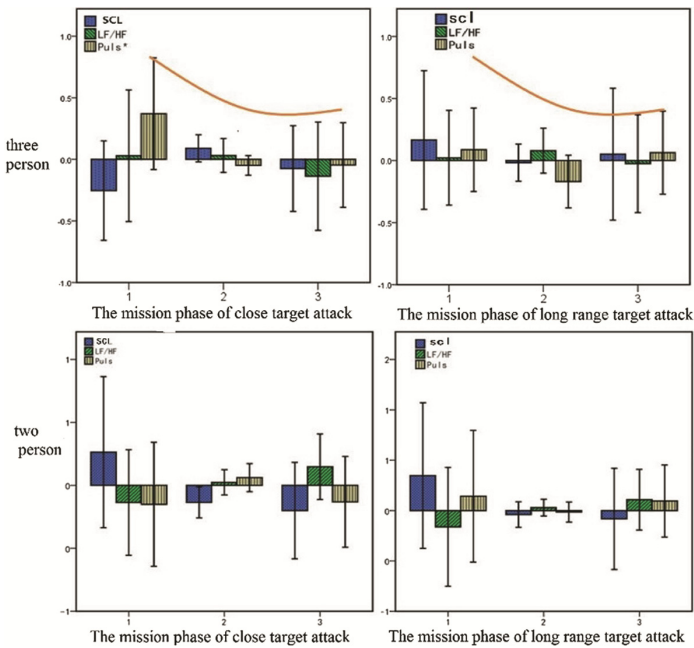


Fig. 4. The normalization of commander physical indexes

The Change Law of the Subjective Score of Multichannel Load in Task Process.

The subjective rate obtains not only the total load of every phase in the tasks, but the visual, auditory, cognitive, behavioral and oral load. Under the different tasks, the summarize of relevancy between the score of total load and each channel is shown in Table 5. The relevancy identified by Pearson Correlation. As a result, there has a reasonably high correlation between the score of task load and the score of visual, cognitive and behavioral channels, particularly in the long range target attack task type. Above

mentioned correlation increase when the executor from gunner shifted to commander, which the score of behavioral channel under the significant influence of the number of occupant ($F(1, 30) = 4.16, p = 0.051$). In close target attack task type, 2-occupant mode has a more obvious decreasing of load score in the cognitive channel than 3-occupant mode.

Table 5. Correlation between task load score of commander and total load score in different channels

| Channel | Occupant's number | | | |
|-----------|-------------------------------|---------|--------------------------|---------|
| | Long range target attack task | | Close target attack task | |
| | 2 | 3 | 2 | 3 |
| Visual | 0.931** | 0.547* | 0.800** | 0.910** |
| Auditory | 0.321 | 0.047 | 0.106 | 0.188 |
| Cognitive | 0.811** | 0.726** | 0.660** | 0.929** |
| Behavior | 0.923** | 0.069 | 0.788** | 0.664** |
| Oral | 0.037 | 0.327 | 0.252 | 0.150 |

Remarks: **means each channel's load score is significantly correlated with the total load score at the $\alpha = 0.01$ level (bilateral)

4 Conclusion

This paper adopts a experimental method which based on human-in-loop simulator to research the task load change because of the change occupants' function. Occupants' task load (in two task types) is evaluated by task performance measurement, subjective estimate method and physiological quantitative measurement technique. We decompose the task types and design subjective evaluation questionnaires according to cognitive channel. We also analyze the result of task performance, subjective evaluation and physiological measurement, and obtain the following conclusions:

The task load of commander in 2-occupant mode is significantly higher than the 3-occupant, but it doesn't decrease the performance of task. On the contrary, the completion time in 2-occupant is less than 3-occupant. The performance of commander is also promoted while in task phase 1 and phase 3. For the long range target attack task type which is more difficult, the time for commander searching target is significantly decreased.

The impact of occupant number change on of subjective evaluation of commander's overall task load is mainly reflected on the task type of long range target strike task in 3-occupant mode, the subjective load score is high in the middle and low in both ends and long distance target is difficult to strike, as a result, the score of the attacking task performance is significantly higher than the other phase. In 2-occupant and 3-occupant mode, the type for close range target strike, are performed by commander attack task. Therefore, in different number of occupant modes, the change of subjective score tends to be consistent.

The total score of task load is highly related to vision, cognition and behavior channel rate. The correlation is increasing when the executor from gunner shifted to commander, particularly in the long range target attack task type. The commander's channel load

score in 2-occupant mode is significantly higher than the 3-occupant mode. Compared the type of close range target task between 2-occupant mode and 3-occupant mode, the score of cognitive channel load decreased significantly in 2-occupant mode.

The conclusion of this study can be made that the change of the occupant number has a greater impact on the occupant performance with a increased physical load within the acceptable range; The evaluation method based on task load is feasible in the semi-physical simulator environment. All in all, the study can provide theoretical support for the evaluation of the occupant assignment function in the new type of PRV. The evaluation method of this study can also be used to other manned operating systems, such as aircraft, ship, and spacecraft and so on.

References

1. Stevena, M.: Human performance and control of multiple systems. *Hum. Fact.* **38**(2), 323–329 (1996)
2. Wang, J., Fang, W., Li, G.: Mental workload evaluation method based on multi-resource theory model. *J. Beijing Jiao tong Univ.* **34**(6), 107–110 (2010)
3. Wickens, C.D., Lee, J., Yili, L., et al.: *Introduction to Human Factors Engineering*, 2nd edn. East China Normal University Press, Shanghai (2007)
4. Cullen, L., Valida, T.: *A Methodology for Predicting Performance and Workload*. Euro Control Experimental Centre, Brussels (1999)
5. Xie, F., Wang, Q., Jin, X., Liao, Y., Zheng, S., Li, L., Zhou, Q., Liu, Z.: Evaluation of the crew workload to quantify typical mission profile special vehicles. In: Long, S., Dhillon, B.S. (eds.) *Man-Machine-Environment System Engineering*. LNEE, vol. 406, pp. 149–158. Springer, Singapore (2016). https://doi.org/10.1007/978-981-10-2323-1_18
6. Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., Griss, M.: Activity-aware mental stress detection using physiological sensors. In: Gris, M., Yang, G. (eds.) *MobiCASE 2010*. LNICSSITE, vol. 76, pp. 282–301. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29336-8_16



Effect of Fatigue and Nervousness of Tower Controller on the Control Efficiency

Xingjian Zhang^(✉), Peng Bai, Xinglong Wang, and Yifei Zhao

Civil Aviation University of China, Tianjin 300300, People's Republic of China
jzhyoung@163.com, pbai@cauc.edu.cn, xinglong1979@163.com,
yifei6666@sina.com

Abstract. Fatigue and nervousness are two of the most common bad states of air traffic controllers at work. To analyze the effects of fatigue and nervousness of tower controllers on the control efficiency, a tower control simulation experiment was designed to collect 22 participants' control performance data and the data of 19 participants were collected successfully. Four states, sober (SO), fatigue (FA), nervous (NE) and fatigue & nervous (FN), were designed. Seven indices of control efficiency were defined and analyzed, including three objective indices: task duration (TD), mean called frequency (MCF) by pilot per-flight and mean speech service time (MSST) for each flight and four subjective evaluation indices: instruction moment score (IMS), situation awareness score (SAS), transient mistake times (TMT) and result mistake times (RMT). The analysis results showed that all the indices were significantly different among the four states. Both fatigue and nervousness can impair control performance and reduce control efficiency. Under the influence of fatigue or nervousness, controllers' initiative and situation awareness will decrease and be more likely to make transient mistakes. At same time, controllers in fatigue state need more time and more communication speech per flight to manage the operation. Controllers in nervous state will make mistake more easily than sober and fatigue states. It can be inferred that fatigue mainly makes controllers' work speed slower and nervousness leads to more control mistakes. These findings are expected to improve the optimization of control efficiency and work management of air traffic.

Keywords: Tower controller · Control efficiency · Fatigue · Nervousness

1 Introduction

Air traffic controller, responsible to provide control services for flight, is one of the key units in air traffic service system. Controllers' working performance has crucial importance for air traffic operational efficiency and air traffic safety. A survey showed that about 52% air traffic management incidents were related with human error of controllers [1]. Maintaining good physical condition is the precondition of high working performance. However, there are some bad physical or mental states which may impair controllers' working performance and reduce control efficiency and safety. Fatigue and nervousness are two of the most common bad states for controllers at work. A report about the relation between human factor and incidents showed that 46.07% controllers

admitted that they have the experience of fatigue duty and had high risk to result in incident [2]. The relation between air traffic safety and the bad states has been studied by some researchers. But it is not clear about the relationship between the efficiency and bad states of controllers. It is essential to explore the effect of the bad states on the efficiency to make better regulation policy for air traffic controllers and improve control efficiency.

There are very few studies focusing on the effect of fatigue and nervousness on control efficiency. Some researchers summarized that the bad states such as fatigue and nervousness will significantly impair the control performance. Controllers' ability will decrease under the influence of the states and some management policy need to be made [1–5]. Additionally, the characteristics of air traffic controller were explored. A study report indicated that fatigue can reduce controllers' ability to carry out a task significantly [6]. Some researchers also studied the detection method in real time based on controllers' facial expressions or work time [7–9]. We can get some fatigue characteristics of air traffic controller like eye movement from these studies. Much less of researches focused on the features of controllers' nervous state. However, these studies only stated the negative effect. The affecting aspect details of fatigue and nervousness on control efficiency are not clear.

Considering the lack of the effect characteristics of fatigue and nervousness on control efficiency, we tried to explore the characteristics in this study. The flight operational efficiency in airport, related with tower controllers' working performance, is the key point for flight rate and overall efficiency. Therefore, the purpose of this study was to analyze the effects of fatigue and nervousness of tower controllers on control efficiency and to analyze the effect differences of the two states. A tower control simulation experiment was designed to collect control data and some indices of control efficiency were defined and analyzed. The results are expected to provide useful references for work management and on-duty arrangement of controller team.

2 Method

2.1 Equipment

In order to collect control efficiency data under the influence of fatigue or nervousness without confounding factors on participants and considering the air traffic safety, a tower control simulator was used in this experiment. The simulator can provide virtual scenario of Qingdao Airport, as shown in Fig. 1. It consists of two radio communication microphones and seven network computers, which provide control system, approach radar, airport surveillance radar, virtual captain and three scenario display respectively. The airport scenario was projected onto three large screens in front, providing a 120° field of view. During the experiment, flights were designed in the control system firstly. Air traffic controller gave control commands to virtual captain through radio with assistance of the radar and airport view. Then, the captain control the flights to fly according to the commands. The speech of all controller in the experiment was recorded with 44100 Hz by a voice recorder, whose type is Sony PCM-D100.

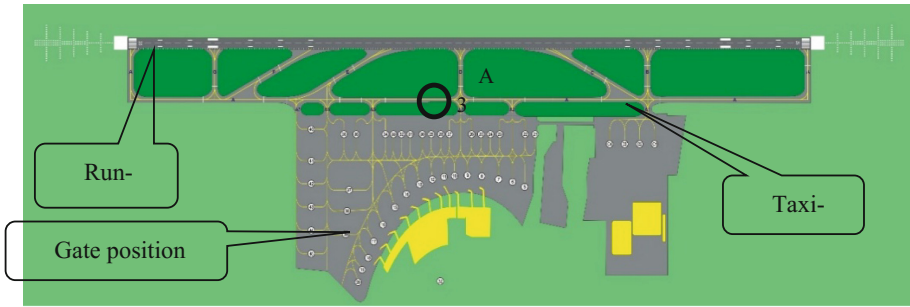


Fig. 1. Simulate airport in the experiment

2.2 Participants

Young air traffic controllers are the major part of frontline staff and compared with older controllers, they may lead to worse results under the influence of fatigue because of less work experience. At the same time, the novice may be nervous easily in control work. Thus, studying young controllers who work with fatigue or nervousness is crucial to improving air traffic safety. Furthermore, almost all the controllers in China are male. Therefore, 22 healthy, young male subjects were recruited to participate in the research. The average age was 24 (SD = 0.52, range = 23–26 years). All participants have possessed a valid controller's license and worked for 1 to 2 years (Avg. = 1.3). A regular circadian rhythm and no drug use were included in the recruiting criteria. We investigated all subjects' sleep rhythm and drug use before recruitment. During each simulated control experiment, they were investigated drug use with a questionnaire. Only subjects with no drug use and no other physical change were allowed to do the experiment. Before recruitment, participants were required to have a test to be familiar with the control simulator and make sure they could perform task as well as the real circumstance. In addition, 4 veterans who are familiar with the flight running rules participated in the experiment to be virtual pilots. All participants agreed and signed an informed consent before participating in the study, and they were paid for their participation.

2.3 Experiment Design

To assess the effects of fatigue and nervousness on the control efficiency, participants were required to conduct experiments at four different states, sober (SO), fatigue (FA), nervous (NE) and fatigue & nervous (FN) states. The group in sober state was considered the control group and the groups in other states were treatment groups. In the experimental design, the key point is to induce fatigue and nervous state for participants. The fatigue state was designed mainly through experiment time according to the drivers' sleep rhythm and the nervous state through control task. The recruitment survey showed that they generally slept from 11 pm to 7 am of next day, had a noon break between 11:30 am to 1:00 pm and begin working again at 1:30 pm. Therefore, the SO and NE state experiments were executed at 9–11 am or 3–5 pm and they were required to sleep

well at least three days before experiments to avoid the effect of fatigue. The FA and FN states experiments were designed to be carried out at 1 am to 3 am, when the participants would easily become fatigue. Also, the participants were required to get up before 8 am on the previous day and not to sleep before the experiment. At the same time, the control tasks in sober and fatigue states were designed as 6 departure flights and 6 arrival flights in about 30 min, which means that the participants can complete the task easily. In NE and FN states, the traffic flow were designed as 10 departure flights and 10 arrival flights in about 35 min and the taxiway of A3 was closed (shown in Fig. 1), which is a complex task to the participants to induce nervousness. The participants were asked to refrain from having any stimulating food or beverage, such as alcohol and drugs.

A questionnaire was used in the experiment to collect data on subjective fatigue and nervous degree. The fatigue degree was set to 7 levels: (1) active, alert, or wide awake; (2) functioning at high levels but not at peak or unable to concentrate; (3) somewhat foggy or let down; (4) foggy, losing interest in remaining awake or slowed down; (5) sleepy, woozy, or prefer to lie down; (6) sleep onset soon or having dream-like thoughts; and (7) asleep. The nervous degree was also designed to 7 levels: (1) very calm, accomplish the control task with ease; (2) calm and complete the task smoothly; (3) a little nervous sometimes; (4) nervous slightly and be not able to judge recollectedly; (5) nervous but be able to accomplish the task; (6) very nervous, only complete part task; (7) nervous very much and be not able to work.

Each participant was required to carry out the four states experiment in four separate days and the interval was at least 3 days. The order of the four states experiment for each participant was random. The procedure of each experiment was as follow:

- First, a participant was instructed regarding the operation of the simulator and the tasks in experiment. Then he was asked to get ready flight progress strip and be familiar with the simulator.
- Second, the questionnaire about subjective fatigue and physical state was completed. Only participant who met the designed criterion can carry out the experiment.
- Third, he was required to try his best to complete well the air traffic control experiment task. The air-ground communication speech, control mistakes and evaluation of control performance were recorded.
- Finally, at the end of experiment, the participant was asked to fill out the questionnaire survey to collect his subjective fatigue and nervous level in experiment.

In the experiment, the control tasks in SO and FA states were similar but different, as well as NE and FN states. In the process of experiment, every virtual captain was required to communicate with controllers with the same criteria to reduce interference on speech data.

2.4 Data Collection and Analysis

The simulated air traffic control data of 19 participants in the four states were collected successfully in the experiment. The participants were asked the fatigue level twice, before and after experiment. The two fatigue levels were used to evaluate the fatigue state during the experiment process. The mean subjective fatigue level of all participants

in the experiment was shown in Fig. 2 and the nervous level in Fig. 3. It indicated that the fatigue level in the designed fatigue state was obviously higher than SO and NE states and it met our expectation, so was the nervous level.

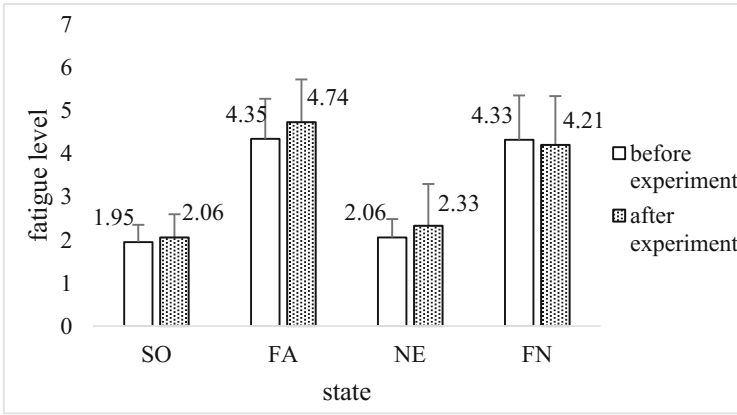


Fig. 2. Mean subjective fatigue level of all participants in the experiment

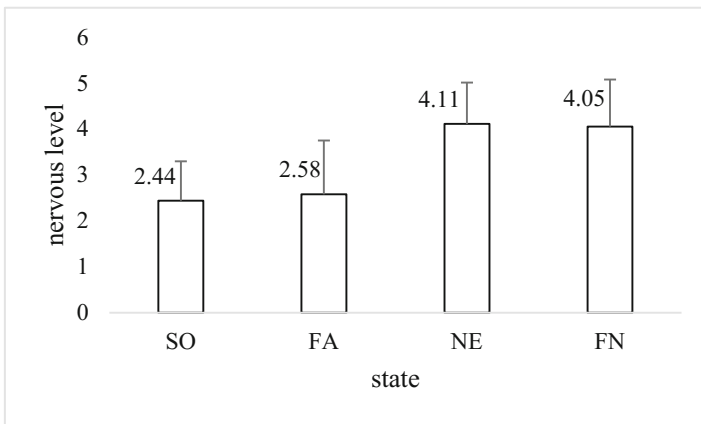


Fig. 3. Mean subjective nervous level of all participants in the experiment

In this study, the main purpose was to obtain the effect characteristics of fatigue and nervousness on control efficiency. Some objective and subjective indices were defined and calculated.

Firstly, three objective indices: task duration (TD), mean called frequency (MCF) by pilot per-flight and mean speech service time (MSST) for each flight were defined. TD was the whole duration of each control experiment, represented the overall working speed. Considering the similar control task in SO and FA states as well as NE and FN states, TD can be only used to analyze the effect of fatigue on control speed. MCF was defined as that divide the whole times of controller called by pilot in the air-ground

communication speech by flight number. It can be used to explain the control initiative and efficiency. MSST was the average speech time duration that controller spent on each flight, representing the efficiency of control speech.

Secondly, four subjective evaluation indices: instruction moment score (IMS), situation awareness score (SAS), transient mistake times (TMT) and result mistake times (RMT) were calculated and extracted. During the experiment process, the control performance was evaluated and the indices were calculated based on the evaluation. IMS meant the timeliness of control instruction, with great importance for control efficiency. In calculation, once participant did not give instruction in appropriate time, certain score would be deducted according to detail situation. SAS was calculated with the same way. TMT meant the transient mistake such as instruction error, call-sign error and flight progress error of controller in once experiment. RMT was the occurrence number of unexpected situation like flight delay and go-around during each experiment. Both the mistake times were related closely with control efficiency.

To evaluate the effect of fatigue and nervousness on the above indices, ANOVA with repeated measures and contrast analysis were used to study the differences in different states. Each index was analyzed with ANOVA only considering one factor, state with four levels firstly. The interaction effect of fatigue and nervousness was also analyzed and discussed. In contrast analysis, we mainly focused on the contrast of FA VS SO state, NE VS SO state, FN VS FA state and FN VS NE state to discuss the effect of fatigue and nervousness on control efficiency.

3 Result

3.1 Effect of Fatigue and Nervousness on Objective Indices

The three objective indices were calculated based on the air-ground communication speech data, which was the main work pattern. The three indices from the four states were analyzed respectively. The results showed that they were all significantly different among the four states. The statistics analysis results of the three indices were shown in Table 1.

Table 1. Statistics analysis results for the three objective indices

| Index | F | P | FA VS SO | NE VS SO | FN VS FA | FN VS NE |
|-------|--------|-------|----------|----------|----------|----------|
| TD | 145.35 | <.001 | .016 | \ | \ | .001 |
| MCF | 8.15 | .001 | <.001 | .048 | .223 | .010 |
| MSST | 140.91 | <.001 | .018 | <.001 | <.001 | <.001 |

TD was obviously different in the four states and the means of it were shown in Fig. 4. No interaction effect of fatigue and nervousness was found for this index. The contrast analysis indicated that it was significantly longer in FA state (Avg. = 35.36, SD = 4.14) than SO state (Avg. = 32.56, SD = 2.84) and longer in FN state (Avg. = 47.62, SD = 4.51) than NE state (Avg. = 42.70, SD = 4.57). It meant that controllers needed

more time to accomplish similar control task under the influence of fatigue, indicating that fatigue might impair the control speed.

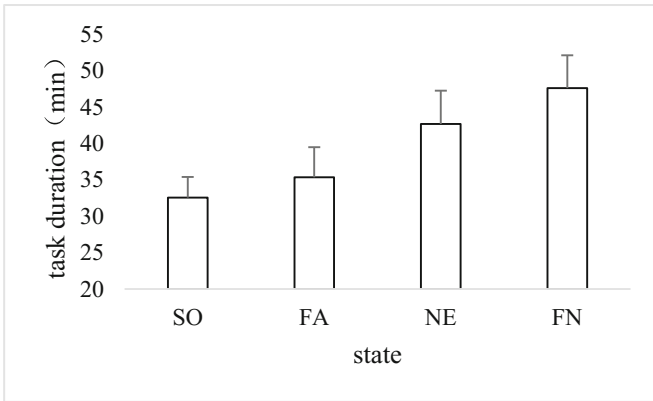


Fig. 4. Means of TD in the four states

The ANOVA analysis showed that MCF was significantly related with controllers’ state and the means of it were shown in Fig. 5. No interaction effect of fatigue and nervousness was found for MCF. The contrast analysis showed that it was significantly more in FA (Avg. = 2.78, SD = 0.61) state and NE state (Avg. = 2.60, SD = 0.47) than SO state (Avg. = 2.37, SD = 0.44) and more in FN state (Avg. = 2.96, SD = 0.70) than NE state. MCF represented the initiative and control efficiency. The increase of it means the decrease of controllers’ control ability and then he will be called by pilot more times to complete the task. The results showed that both fatigue and nervousness could impair control ability and efficiency. The significant difference between FN and NE but not

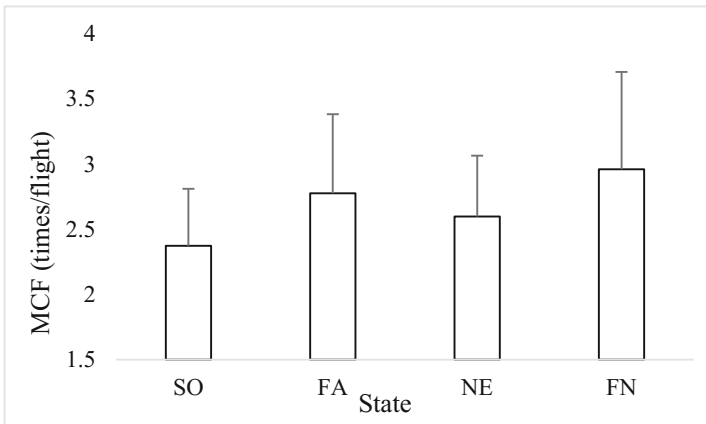


Fig. 5. Means of MCF in the four states

between FN and FA may indicate that fatigue has a more important role than nervousness in the effect.

MSST was significantly affected by state and the means of it were shown in Fig. 6. There was significantly interaction effect of fatigue and nervousness on MSST ($p < .001$). The contrast analysis showed that it was significantly different between any two states. MSST was longer in FA (Avg. = 38.51, SD = 3.60) state than SO state (Avg. = 37.93, SD = 3.37), shorter in NE (Avg. = 34.69, SD = 2.17) state than SO state and much longer in FN (Avg. = 45.60, SD = 3.77) state than FA and NE state. The results indicated that under the influence of fatigue, controllers need longer speech time to control each flight. However, the time became shorter when they were nervous. At same time, in FN state, controllers needed much longer speech time than any other states, which was due to the interaction effect of fatigue and nervousness.

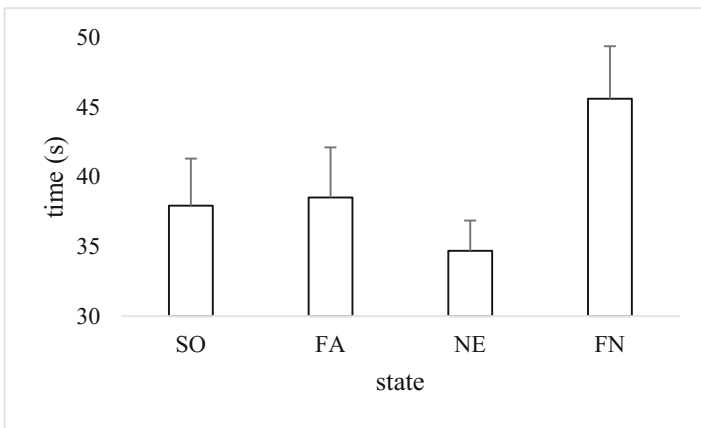


Fig. 6. Means of MSST in the four states

3.2 Effect of Fatigue and Nervousness on Subjective Indices

The four subjective indices were calculated based on the evaluation data for control process. The data was recorded by a veteran controller and evaluated according to a certain criterion. The four indices from the four states were also analyzed respectively. The results showed that they were all significantly different among the four states and no interaction effect of fatigue and nervousness was found for the indices. The statistics analysis results of the three indices were shown in Table 2.

Table 2. Statistics analysis results for the three objective indices

| Index | F | P | FA VS SO | NE VS SO | FN VS FA | FN VS NE |
|-------|-------|-------|----------|----------|----------|----------|
| IMS | 10.14 | <.001 | <.001 | .004 | .016 | .069 |
| SAS | 19.77 | <.001 | .018 | <.001 | .001 | .662 |
| TMT | 20.04 | <.001 | .001 | <.001 | <.001 | .106 |
| RMT | 12.96 | .003 | .120 | .007 | .023 | .277 |

The ANOVA analysis results showed that IMS was significantly affected by controllers' state and the means of it were shown in Fig. 7. The contrast analysis revealed that it was significantly lower in FA (Avg. = -5.32, SD = 3.15) state and NE state (Avg. = -6.21, SD = 4.66) than SO state (Avg. = -2.47, SD = 2.57) and lower in FN state (Avg. = -9.42, SD = 5.91) than FA state. IMS represents the ability of giving control commands at the right moment, which is important for control efficiency. The analysis results showed that both fatigue and nervousness would affect this ability. The obviously differences between FN and FA state but not between FN and NE may reveal that nervousness impairs the ability of instruction moment more seriously.

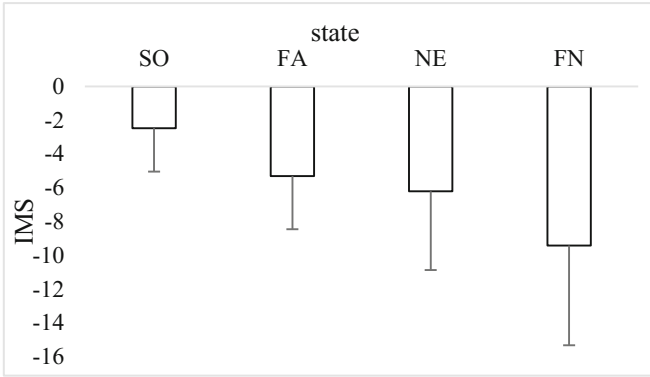


Fig. 7. Means of IMS in the four states

The analysis results of SAS indicated that it was also significantly related with controllers' state, as shown in Fig. 8. The contrast analysis revealed that it was significantly lower in FA (Avg. = -5.16, SD = 4.32) state and NE state (Avg. = -10.89, SD = 5.38) than SO state (Avg. = -2.26, SD = 2.25) and lower in FN state (Avg. = -11.63, SD = 4.27) than FA state. SAS represents the comprehensive air traffic control ability in the working process. Situation awareness is crucial for controllers to have correct

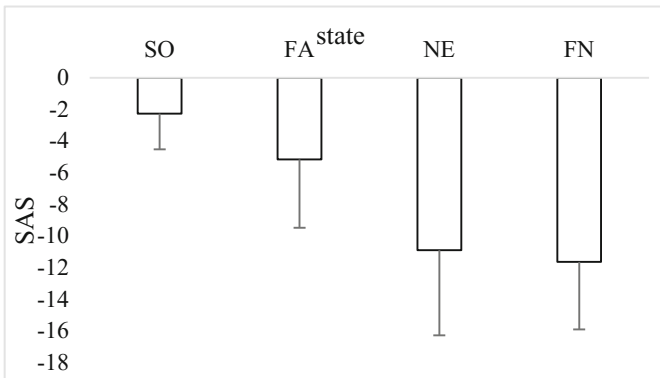


Fig. 8. Means of SAS in the four states

judgment and give appropriate control commands. The results showed that the ability would be impaired by fatigue and nervousness. The comparison may also reveal that nervousness affects the ability of situation awareness more seriously than fatigue.

TMT was significantly affected by state and the means of it were shown in Fig. 9. The contrast analysis showed that it was significantly more in FA (Avg. = 10.84, SD = 4.54) state and NE state (Avg. = 16.58, SD = 7.22) than SO state (Avg. = 7.26, SD = 3.03) and more in FN state (Avg. = 20.16, SD = 10.07) than FA state. TMT is the statistics for controllers' control errors, representing the characteristics of control process. The analysis results revealed that under the influence of fatigue or nervousness, controllers would make more control mistakes. The obviously differences between FN and FA state but not between FN and NE may still indicate that nervousness will lead to more errors than fatigue.

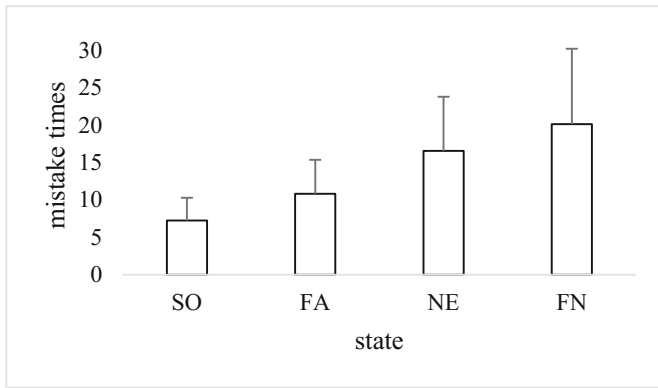


Fig. 9. Means of TMT in the four states

The analysis results showed that RMT was significantly related with state, as shown in Fig. 10. It was significantly more in NE state (Avg. = 2.26, SD = 1.69) than SO state

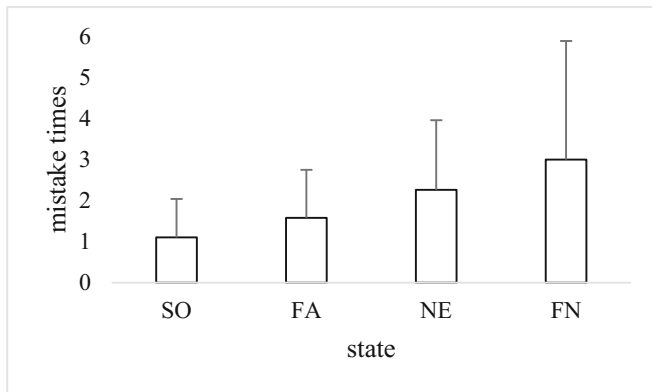


Fig. 10. Means of RMT in the four states

(Avg. = 1.11, SD = 0.94) and more in FN state (Avg. = 3.00, SD = 2.8) than FA state (Avg. = 1.58, SD = 1.17). RMT is the evaluation of control performance, representing the characteristics of control result. The analysis results indicated that nervousness would generate more bad traffic situations. The comparison between FN and FA state also proved the significant effect of nervousness on control result mistake.

4 Discussion

The analysis results indicate that all the indices are significantly different in different states. Under the influence of fatigue, controllers' control speed will be slow, initiative decrease, control ability on instruction moment and situation awareness decrease and control mistake increase. When working in nervous state, controllers' control speed will be slow and initiative decrease slightly, but make much more transient and result mistake. Fatigue and nervousness have significant interaction effect on the speech service time for each flight.

For the effect of fatigue, it may be easily understood that the control speed become slower in fatigue state. It has been proved that people's somatic function will become slower [10]. The controllers' ability of perception, reaction and action will weaken when fatigue. All of these changes characteristics will lead to slower of controllers' action, including decision-making and speech speed. Considering these change characteristics, we can deduce that controllers' mentality will be lazy, which make their initiative decrease. Because of these reasons, controllers will need more time and more speech time to complete task and be called more times by each pilot. At the same time, the decrease of control ability on instruction moment and situation awareness is also due to the impairment of fatigue on somatic function. Then, the decrease of ability results in the increase of control mistakes. It can be summarized that fatigue impairs comprehensive control ability and mainly decrease processing speed and mentality initiative. Control efficiency will be significantly decline under the influence of fatigue.

For the effect of nervousness, the mainly reason may be the complicated and heavy control task for controllers. The higher MCF may be due to that controllers cannot deal with every flight immediately because of vast control task from much more flight. In nervous experiment, due to that controllers need to accomplish plenty of control task in certain time, the speech speed might become faster and then lead to lower mean speech service time. Similarly, the control task was very complex in the situation of high traffic flow, making nervous state for controllers. They could not give appropriate commands to each flight timely and hold the running situation correctly under the influence of nervousness. Then more control mistake appeared during the control process. In a word, heavy control task is easy to cause nervous state and nervousness affects control efficiency mainly through control performance. Much more mistakes may appear under the influence of nervousness.

The interaction effect of the two bad states only appeared in the index of MSST. It could be deduce that fatigue impaired control ability firstly and in the FN state experiment, controllers ability could not met the need of the complicated task. Then they had to spend much more speech time to complete the task. At the same time, it can found

that under influence of both fatigue and nervousness, every index showed the worst result. We can conclude that the combination of more bad states will impair control performance and efficiency much more seriously.

The results of this study showed some characteristics effects of fatigue and nervousness on control efficiency. The results have potential application in practical use. They are helpful to have a better understanding on the effect characteristics of fatigue and nervousness on control ability and make countermeasure. The results also provided a reference for further study on the effect characteristics of controllers' state. In further study, more indices of control performance need to be analyzed and some countermeasures will be studied to prevent the impairment on control efficiency.

5 Conclusion

This paper explores the effect of controllers' fatigue and nervousness on control efficiency, 7 indices were defined and analyzed. They can explain different aspects of control efficiency. Based on the analysis results, the following conclusions can be made:

- Fatigue can make controllers' control speed be slower and control initiative decrease. Their control ability on instruction moment and situation awareness will be impaired by fatigue and more transient mistake will appear in control process.
- Complicated control tasks are easy to lead to controllers' nervous state. Under the influence of nervousness, controllers' control speed and initiative will decrease. Nervousness will result in much more transient and result mistake.
- Controllers need much longer speech time per flight to complete task because of the interaction effect of fatigue and nervousness. Under the influence of the two states simultaneously, the control performance will become much worse.
- Comparing the effect characteristics of the two state, fatigue mainly makes controllers' work speed and initiative decrease due to decline of somatic function. Nervousness mainly leads to more control mistakes because of the intricate task.

In summary, both fatigue and nervousness will impair air traffic controllers' work ability and control efficiency. These effect characteristics are expected to provide reference for the work management of air traffic controller.

Acknowledgments. This study was supported by the National Natural Science Foundation of China project: Research on the Recognition Method of Bad working state of Controller Based on Individual Speech Characteristics, No. U1533117, the National Key Research and Development Plan of China project: Tracing, Recognition and Warning of High Risk Flight Trajectory, No. 2016YFB0502405, and the Science Research Starting Foundation of Civil Aviation University of China (No. 2014QD02X).

References

1. Zhang, H.H.: Human factors in air traffic control safety. *Manag. J.* **18**, 130 (2010)
2. Song, H.M.: Effect of controllers' psychology factors on safe. *Technol. Innov. Appl.* **22**, 318–319 (2012)
3. Xu, H.N.: Effect of controllers' psychology factors on safety. *Ind. Sci. Trib.* **13**, 141–142 (2014)
4. Huang, J.L.: Effect of controllers' bad psychology on safety and solutions. *Air Traffic Manag.* **12**, 37–40 (2007)
5. Wu, P., Mei, X.: Essay on the effect of controllers' mental quality for security of control and the strategies to improve. *J. Civ. Aviat. Flight Univ. China* **25**, 70–73 (2014)
6. CRATCOH: Report of a Committee on Regulation of Air Traffic Controllers' Hours to the Civil Aviation Authority. Civil Aviation Authority, Cheltenham (1990)
7. Sun, R.S., Shi, Z.P., Wang, J.H.: Face recognition for fatigue risk assessment of air traffic controllers. *J. Transp. Inf. Saf.* **32**, 1–4 (2014)
8. Lei, W., Ruishan, S.: Analysis on flight fatigue risk and the systematic solution. In: Robertson, M.M. (ed.) *EHAWC 2011*. LNCS, vol. 6779, pp. 88–96. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21716-6_10
9. Gander, P., Hartley, L., Powell, D., et al.: Fatigue risk management: organizational factors at the regulatory and industry/company level. *Accid. Anal. Prev.* **43**, 573–590 (2011)
10. Zhang, X.J., Zhao, X.H., Du, H.J., Rong, J.: A study on the effects of fatigue driving and drunk driving on drivers' physical characteristics. *Traffic Inj. Prev.* **15**, 801–808 (2014)

Situation Awareness, Training and Team Working



Dynamic Prediction Model of Situation Awareness in Flight Simulation

Chuanyan Feng¹, Xiaoru Wanyan^{1(✉)}, Shuang Liu², Damin Zhuang¹,
and Xu Wu³

¹ School of Aeronautics Science and Engineering, Beihang University,
No. 37 Xueyuan Road, Haidian District, Beijing 100191, China
fengchui Fengfly@qq.com,

{wanyanxiaoru, dmzhuang}@buaa.edu.cn

² Marine Human Factors Engineering Lab,
Institute of Marine Technology and Economy, Beijing 10008, China
liushuangbh@163.com

³ AVIC China Aero Polytechnology Establishment, Beijing, China
e126126_19@126.com

Abstract. Dynamic prediction for pilot situation awareness (SA) is an important issue in aviation safety. This paper presents a dynamic prediction model on the basis of the progressive triggering relationship between low and high-level SA. Six typical cognitive status (“Unnoticed”, “Attention of situation element (SE) but not reaching perception”, “Perception of SE”, “Perception but not matching the best rule”, “Triggering of the best rule” and “Decision making and operation”) were proposed for the description of the cognitive process of SE. Eighteen participants were selected to conduct the flight simulation tasks, and the situation awareness global assessment technique (SAGAT) method was adopted to measure the performance data (including accuracy and response time) at 13 typical time points. Statistical analysis showed that the theoretical value of the proposed SA dynamic prediction model was significantly correlated with accuracy and response time, which validated the model in the flight simulation environment preliminary. The proposed SA dynamic model in flight scenarios can give some references for cockpit’s human-computer interface design and flight tasks optimization assignment.

Keywords: Situation awareness · Mathematical model · Flight simulation
Cognitive assessment · Interface design

1 Introduction

Situation awareness (SA) is closely related to aviation safety. The concept of SA firstly appeared in aviation psychology [1]. Since the late 1980s, there has been a growing number of SA-related studies and has drawn considerable attention from academics [2]. After more than 20 years of research and exploration, SA is now one of the most important studies in ergonomics [2–4]. SA is not only related to interface design that supports SA, but also to disasters and accidents that lack of SA, especially for dynamic and safety-centric operating environments. The statistical result of aviation accidents

revealed that 35.1% non-major accidents and 51.6% major accidents were caused by the failure of pilot's decision-making, and the main reason for that was the lack of SA or SA error instead of the error in decision-making. SA can be easily lost when the changing situation was not fully understood by flight crew, which may lead to the controlled flight into terrain (CFIT), such as the Air France Flight 447 [2]. In air traffic control tasks, air traffic controllers (ATC) need to keep abreast of the current aircraft dynamics, or else they may have disastrous consequences, and the American Airlines Flight 1493 in 1991 is a typical example [5]. In the field of nuclear power, the maintaining of a good SA is also crucial for operator's safety operation, especially in emergencies, such as the Three Island Mile accident in 1979 [6].

The concept of SA has been controversial, with more than 30 definitions [7], of which Endsley's view is more widely accepted. She pointed out that SA includes three levels of perception, comprehension and prediction, namely "perception of environmental components in a large amount of time and space, understanding of its meaning and prediction of the status in the near future" [1]. At present, studies on SA mainly focus on the theoretical models and measurement methods. The construction of the SA theoretical model is one of the current research difficulties [2, 7, 8]. The study of SA mechanism models is an explanation and extension of the definition of SA, which relies significantly on cognitive psychology and gives explanations of the SA formation process in individual brain [2]. The three-level model, perceptual cycle model [9] and theory of activity model [10] are currently recognized as the three main SA mechanism model.

Since the concept of SA emerged, it has also been one of the researchers' concerns to establish a quantitative computing model for SA [8, 11, 12]. Up to now, researchers have carried out a series of researches on the quantitative calculation model of SA. For example, Wickens et al. established the Attention-Situation Awareness (A-SA) based on the theory of attention distribution to predict performance errors of pilot [11]. Kirlik and Strauss constructed the SA ecological model by assigning ecological validity to the SE (Situation Element) [12]. Entin's PSM (Performance Sensitivity Model) emphasizes the dynamics of SE and uses sensitive coefficients to reflect the impact of SE on SA [13]. The SA level in the MIDAS (Human Machine Integration Design and Analysis human performance Model) model of Hooey et al. is calculated by the ratio of the actual state SA level to the ideal state SA level. Liu et al. put forward the SA model based on attention resource allocation [15] and ACT-R (Adaptive Control of Thought Rational) cognitive theory, which conducted the prediction of SA level in flight instrument display scenarios [8].

This paper presents a dynamic prediction model that six typical cognitive status ("Unnoticed", "Attention of SE but not reaching perception", "Perception of SE", "Perception but not matching the best rule", "Triggering of the best rule" and "Decision making and operation") were used for the description of the cognitive process of SE. A flight simulation experiment was conducted among eighteen participants, and the statistical analysis was used for the validation of SA model. The proposed SA quantification model of individual can be used to assess and improve SA of the current designs, to predict situations where SA losses may occur, and to improve operator performance [11, 16]. In addition, in a typical aviation environment, the individual SA model can be further extended to the assessment of team SA (consisting of flight crews, air traffic controllers, etc.) and system SA (consisting of flight instrumentation, autopilot, etc.) [2].

2 Theoretical Modeling

Assume that there are n SEs (Situation Elements) in the situation at time t , and the operator's cognitive level to SE_i is $P_i(t)$. $P_i(t)$ stands for operator's knowledge of SE_i at present and in the future, and a higher $P_i(t)$ indicates a better understanding of the information in the current and future status.

$$SA(t) = f(P_1(t), P_2(t) \dots P_i(t), P_n(t)) \quad (1)$$

Here $SA(t)$ is the SA level of operator at time t , which is closely related to each relevant SEs . Note that the true value of SE_i at time t is $O_i(t)$, the characterization value in operator's brain is $S_i(t)$, and then the $P_i(t)$ can be represented by $S_i(t)$, $O_i(t)$ and uncertain errors $x_i(t)$ [14], see Eq. 2.

$$P_i(t) = f(S_i(t), O_i(t), x_i(t)) \quad (2)$$

$S_i(t)$ is related to characterization value $S_i(t-1)$ and action $g_i(t-1)$ in previous time, as well as several internal factors $k_i(t)$ (operator's memory and knowledge) and external factors $d_i(t)$ (physical display characteristics of SE [13]), therefore

$$S_i(t) = f(S_i(t-1), g_i(t-1), k_i(t), d_i(t), x_i(t)) \quad (3)$$

Here the external factors $d_i(t)$ work through internal mechanism $k_i(t)$, and thus affect its characterization state [1].

$$d_i(t) = f(k_i(t), x_i(t)) \quad (4)$$

By considering the progressive trigger relationship [1] [8] in cognitive activity, only low-level cognitive activities accumulated to a certain amount can they enter into the next phase of high-level activities, and cause changes in the quality of cognitive level. Therefore, the level of cognition to a certain SE can be regarded as a discrete value changing with the cognitive stages.

As shown in Fig. 1, the cognitive levels were set as follows: (1) Attentional behavior of SE_i did not occur (event \overline{ai}); (2) Attention behavior occurred but did not reach to perception (event $ai\overline{bi}$); (3) Attention occurred and then reach to perception; (4) Pattern matching but failed to trigger the best rule (event $ai\overline{bici}$); (5) The best rule triggered to form an understanding of current state, that is, to reach the corresponding SA2; Or to form an understanding of the future state, that is, to achieve the corresponding SA3 (event $aibici$); (6) Decision-making and operation.

The cognitive level at time t is determined by the completion state of SE_i in cognitive circuit [14]:

$$P_i(t) = \sum_{j=1}^{SUM_{road}} road_j S_i(t)/O_i(t) \quad (5)$$

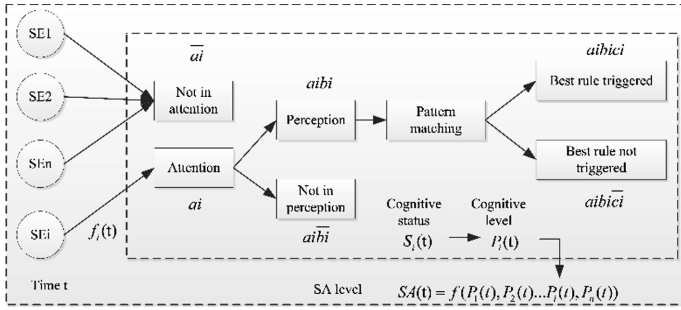


Fig. 1. Representation of SA level in cognitive process

(1) Attention behavior did not occur

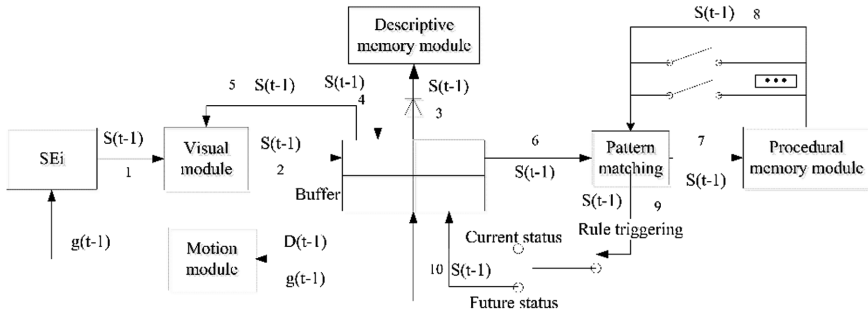


Fig. 2. Description of SA cognitive process

There is only a SE may be noticed at time t , and the individual always tends to choose the most valuable one. The higher the value is, the greater the probability is chosen. Note attention behavior of SE_i as event ai , then

$$p(ai) = f_i(t) \tag{6}$$

Here $f_i(t) = \max(f_1(t), \dots, f_i(t), \dots, f_n(t))$. If SE_i was unnoticed, then

$$p(\bar{ai}) = 1 - f_i(t) \tag{7}$$

There is an updating characteristics in working memory, although theoretically the characterization value of SE_i in operator's brain should be maintained in time $t - \tau$. So,

$$S_i(t) = S_i(t - \tau) \times e^{-k_1 \tau} \quad (8)$$

where k_1 reflects the speed of information updating in working memory of individual. The following $k_1 = 0$ indicates that the SE_i at time $t - 1$ was fully remembered by individual; τ reflects the forgotten time of SE_i , and the unit time was set as 1 below. It can be seen from Fig. 2 that the SE_i has been characterized in a total of 10 cognitive circuits. Since no attention occurs, the cognitive status is not updated at this time, and the memory at time $t - 1$ continues to be maintained. The cognitive status value on each line is as follows:

$$road_j S_i(t) = road_j S_i(t - 1) \times e^{-k_1} \quad (9)$$

At this time, the cognitive level of SE_i exists in the following two situations:

(1) When $\frac{O_i(t) - S_i(t)}{O_i(t) + \epsilon_{min}} = \frac{O_i(t) - e^{-k_1} S_i(t-1)}{O_i(t) + \epsilon_{min}} \leq \Delta$, the difference between $S_i(t)$ in brain at time $t - 1$ and $O_i(t)$ at time t is considered to be within an acceptable error range

Δ . So the cognitive level is $P_i(t) = \sum_{j=1}^{10} road_j \frac{1}{10} \times \frac{S_i(t)}{O_i(t)} = \frac{S_i(t-1) \times e^{-k_1}}{O_i(t)}$. (2) When

$\frac{O_i(t) - S_i(t)}{O_i(t) + \epsilon_{min}} = \frac{O_i(t) - e^{-k_1} S_i(t-1)}{O_i(t) + \epsilon_{min}} > \Delta$, then the difference between $S_i(t)$ in brain at time $t - 1$ and $O_i(t)$ at time t is considered to be beyond an acceptable error range Δ .

(2) Attention but not reach perception

Selective attention occurs at time t , and the visual module obtains the new $S_i(t)$ of SE_i with probability of $f_i(t)$ and passes it to the buffer module. The buffer module accesses the descriptive module to extract the corresponding descriptive knowledge. However, the amount of activation AC_i does not reach the threshold and cannot form a perception, the probability is:

$$p(ai\bar{b}i) = p(ai) \times p(\bar{b}i/ai) = f_i(t) \times p(\bar{b}i/ai) = f_i(t)(1 - p(bi/ai)) \quad (10)$$

Then probability that selectively attention of SE_i may be occurred at time t is:

$$f_i(t) = A_i(t) / \sum_{i=1}^n A_i(t) \quad (11)$$

where $A_i(t)$ is calculated on the basis of multiple-resource theory [17]. In this case, the cognitive status of line 1–2 is $S_i(t) = O_i(t)$, and the value of line 3–10 is not updated, and the memory of the previous time is maintained. At this time, the cognitive status is $S_i(t) = S_i(t - 1) \times e^{-k_1}$. Similarly, two cognitive level of SE_i were existed: $P_i(t) = 0.2 + \sum_{j=3}^{10} \frac{1}{10} \times \frac{S_i(t-1) \times e^{-k_1}}{O_i(t)}$ or $P_i(t) = 0.2$.

(3) Attention and reach to perception

If $AC_i > Lim$, then the perception of $S_i(t)$ is formed, that is, to reach the SA1. The probability is

$$p(aibi) = p(ai)p(bi/ai) = \frac{f_i}{1 + e^{-(AC_i(t)-\tau)/s}} \quad (12)$$

In this case, the cognitive status of line 1–5 is updated and read a new value, 6–10 remained the same as time $t - 1$. The cognitive level of SE_i is $P_i(t) =$

$$0.5 + \sum_{j=6}^{10} \frac{1}{10} \times \frac{S_i(t-1) \times e^{-k_1}}{O_i(t)} \text{ or } P_i(t) = 0.5.$$

(4) Pattern matching

The buffer passes the value of new status to the procedural memory module, and to match the corresponding procedural knowledge with varieties of rules: $U_1 = P_1G - C_1$; $U_2 = P_2G - C_2$. However, the best rules have not yet selected and triggered at this point, then the probability of occurrence to this situation is:

$$p(aib\bar{c}i) = p(aibi)p(\bar{c}i/aibi) = p(aibi)(1 - p(ci/aibi)) \quad (13)$$

In this case, line 1–5 is updated, line 6–10 is maintained. Then $P_i(t) =$

$$0.8 + \sum_{j=9}^{10} \frac{1}{10} \times \frac{S_i(t-1) \times e^{-k_1}}{O_i(t)} \text{ or } P_i(t) = 0.8.$$

(5) Triggering of the best rule

When the pattern matches the best rule at this time and is triggered by the rule to form an understanding of $S_i(t)$ at present, SA2 or SA3 is considered to be reached; then

$$p(aibici) = p(aibi)p(ci/aibi) = \frac{f_i}{1 + e^{-(AC_i(t)-\tau)/s}} \frac{e^{U_i/\theta}}{\sum_l^m e^{U_l/\theta}} \quad (14)$$

Now line 1–10 is updated, $P_i(t) = \sum_{j=1}^{10} road_j \frac{1}{10} \times \frac{S_i(t)}{O_i(t)} = 1$.

(6) Decision making and operation

According to the formed SA, instructions $D_i(t)$ related to current situation was transmitted to motion module by pilot, then:

$$D_i(t) = f(D_i(t-1), S_i(t), k_i(t)) \quad (15)$$

Now line 1–10 is updated, $P_i(t) = 1$. According to the decision signals, the motion module makes a certain action to $g_i(t)$, and the action feeds back to the buffer module.

$$g_i(t) = f(g_i(t-1), D(t), S_i(t), k_i(t)) \quad (16)$$

Suppose there are n SEs , where e_i represents the influence coefficient of each SE 's cognitive level to current SA , and its value is related to multiple factors. Research indicated that the sensitivity coefficient is related to the average task load \overline{mw}_k , the presentation interval time of information ΔT_{it} , processing time of information ΔT_{pt} [17].

$$e_i = f(\overline{mw}_i, T_i) = \overline{mw}_i * \Delta T_{it} * \Delta T_{pt} \tag{17}$$

Then the final expression of SA is

$$SA(t) = \sum_{i=1}^n e_i P_i(t) = e_1 P_1(t) + e_2 P_2(t) \dots e_i P_i(t) + e_n P_n(t) \tag{18}$$

3 Experimental Validation of SA Model

Two parts were mainly included in the experimental verification of SA model: design of typical situation and achievement of performance measurement. Interface simulation models and design of experimental scheme under simulated flight conditions were included in typical situation. The experimental data were collected by the Situation Awareness Global Assessment Technique (SAGAT) [1].

3.1 Design of Typical Situation

A high-fidelity simulation flight platform was built based on Flight Gear 3.4.0 software in laboratory environment, which includes the Primary Flight Display (PFD), the Navigation Display (ND) panel, and the Engine Indication and Crew Alerting System (EICAS)), shown as Fig. 3(a)–(c). The flight information display interface was presented in three 17-in. LCD screen (see Fig. 3(d)), with the average screen brightness of 120 cd/m^2 , the resolution of 1280×1024 , the ambient light of 600 lx . The Saitek Yoke civil aviation flight joystick system was used to complete the flight operations.



Fig. 3. Experiment scenario and interface design

3.2 Participants

Eighteen participants (average was 22.6 years) were selected to carry out flight simulation tasks, all of whom were simulated pilot who had a good aviation background from Beihang university. They were both in good health, right handed, and with a normal vision or corrected vision.

3.3 Experimental Design

The flight situation which composed of flight mission and interface display factors was mainly investigated in this experiment. The display area was divided into 3 AOIs, namely, PFD (AOI 1), ND (AOI 2) and EICAS (AOI 3). Each participant needs to complete a traffic-pattern flight, in which the “three-four turning and auto-alignment” phase was chosen to validate the model.

All participants were required to have adequate simulated flight training and the formal testing was commenced after the flight operations and experimental procedures had been fully mastered. The corresponding scores for flight task operations and average task in multiple-resource load [18] were shown in Table 1.

Table 1. Operations in flight situations

| No. | Flight operations | Visual | Auditory | Cognition | Motion | Total scores |
|-----|---|--------|----------|-----------|--------|--------------|
| 1 | Keep the height stable at 3000ft, adjust the heading to 14 on MCP panel | 3.7 | 2.0 | 1.0 | 2.6 | 9.3 |
| 2 | After the heading is stable on ND, activate the horizontal navigation | 4.0 | 1.0 | 1.0 | 2.2 | 8.2 |
| 3 | Observe the display on ND and wait for the aircraft to turn on the runway automatically | 3.7 | 1.0 | 1.0 | 0.0 | 5.7 |

A single-factor within-subject design was used in the validation, and the dynamic changes of SA level in different freezing time points which consisting of task operations and instrument displays in the flight scenarios was the independent variable. During the experiment, the experimental interface froze at different experimental time points and the corresponding freezing questions occurred. The participant needs to make response with the mouse within a given time. The corresponding freezing problems are shown in Table 2. The instrumental importance of each AOI was set according to the flight situation [17], and the information expectation was given in the start of the experiment, showed in Table 2.

Table 2. Freezing questions at different time points

| SA level | No. | Display | Description of question |
|---------------|-------|---------|--|
| Perception | 4 | PFD | What is the airspeed at present? |
| | 5 | PFD | What is the roll angle at present? |
| | 6 | EICAS | What is the status of flap at present? |
| | 9 | PFD | What is the pitch angle at present? |
| | 8, 10 | ND | What is the heading at present? |
| | 12 | EICAS | What is the speed of engine N1 at present? |
| Comprehension | 3, 11 | ND | What is the phase of flight at present? |
| | 7 | PFD | Whether the APP state can be activated or not at present? |
| | 13 | ND | Whether the aircraft has been aligned to runway or not at present? |
| Prediction | 1 | ND | What is the heading after 6 s? |
| | 2 | ND | How many seconds later can you activate the horizontal navigation? |

4 Experimental Results and Analysis

4.1 Theoretical Value of Dynamic SA Model and the Experiment Value

The interface design, attention mobility, information expectation and information value in the three monitored AOIs by experimental interface model were calculated, showed in Table 3 [8].

Table 3. Attention allocation elements in AOIs

| Element | PFD | ND | EICAS |
|-------------------------|--------|--------|--------|
| Information expectation | 0.23 | 0.62 | 0.15 |
| Information value | 0.0761 | 0.2324 | 0.0442 |
| Attention mobilization | 0.3540 | 0.2500 | 0.3960 |
| Interface design | 0.5053 | 0.2529 | 0.2418 |

$P(bi/ai)$, $P(ci/bi)$, $P(aibi)$, $P(aibici)$ were calculated by Eqs. 12 and 14 respectively, and combined with the Eq. 17 to obtain the SA level at each time point, the results were shown in Table 4.

Table 4. Coefficient of sensitivity in flight situations

| Parameter | PFD | ND | EICAS |
|-------------|--------|--------|--------|
| $P(bi/ai)$ | 0.7552 | 0.1682 | 0.7876 |
| $P(ci/bi)$ | 0.8000 | 0.8000 | 0.8000 |
| $P(aibi)$ | 0.1079 | 0.1403 | 0.0182 |
| $P(aibici)$ | 0.0863 | 0.1122 | 0.0146 |

4.2 Correlation Analysis

Based on the validation method of SA model used by Wickens et al. [11], a Pearson correlation analysis was performed between predictive value of SA model and experimental measurement results. The variation of SA predictive value, response time and accuracy in different time points were shown in Figs. 4, 5 and 6.

The statistical results showed that there was a significant moderate correlation between SA predictive value and accuracy in performance measurement ($r = 0.642$, $P = 0.018$) and a significant moderate correlation with SAGAT response ($r = -0.554$, $p = 0.049$), which has verified the validity of the model to a certain extent.

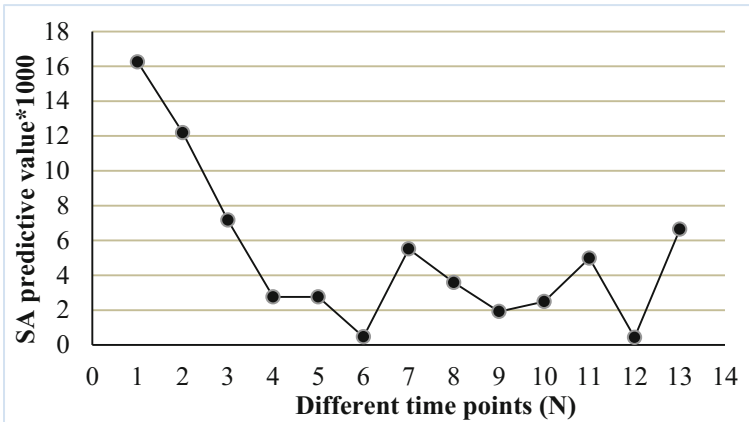


Fig. 4. Variation of SA predictive value in different time points

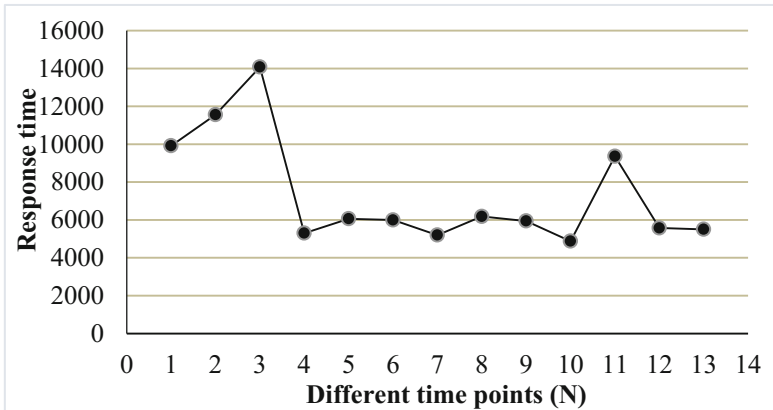


Fig. 5. Variation of response time in different time points

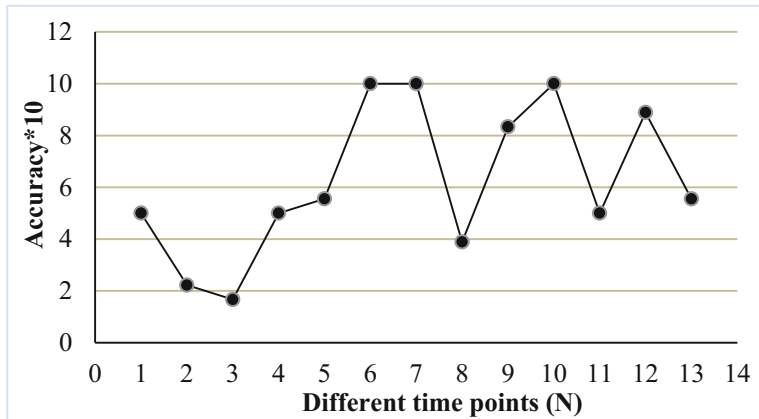


Fig. 6. Variation of accuracy in different time points

5 Conclusion

In conclusion, a new quantitative generation model of SA was proposed in this study. By considering the progressive trigger relationship between low and high-level SA, six discrete values was divided to the cognitive level of a certain information component; the sensitivity coefficient was calculated on multiple-resource theory, and the SA level was then calculated on the basis of conditional probability theory. The verification of SA model is completed on the built simulation platform. Based on the different display instrument, the “three-four turning and auto-alignment” is selected for situation to be analyzed. The verification results showed that the proposed SA model has a certain validity.

Acknowledgement. This study was financially co-supported by the jointly program of National Natural Science Foundation of China and Civil Aviation Administration of China (No. U1733118), as well as the National Natural Science Foundation of China (No. 71301005).

References

1. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. *Hum. Fact.: J. Hum. Fact. Ergon. Soc.* **37**(1), 32–64 (1995)
2. Stanton, N.A., Salmon, P.M., Walker, G.H., Salas, E., Hancock, P.A.: State-of-science: situation awareness in individuals, teams and systems. *Ergonomics* **60**(4), 449–466 (2017)
3. Wickens, C.D.: Situation awareness: review of Mica Endsley’s 1995 articles on situation awareness theory and measurement. *Hum. Fact.* **50**(3), 397–403 (2008)
4. Salmon, P.M., Stanton, N.A.: Situation awareness and safety: contribution or confusion? Situation awareness and safety editorial. *Saf. Sci.* **56**(7), 1–5 (2013)
5. Wickens, C.D., Gutzwiller, R.S., Santamaria, A.: Discrete task switching in overload: a meta-analyses and a model. *Int. J. Hum.-Comput. Stud.* **79**(C), 79–84 (2015)

6. She, M.R., Li, Z.Z.: Design and evaluation of a team mutual awareness toolkit for digital interfaces of nuclear power plant context. *Int. J. Hum.-Comput. Interact.* **33**(9), 1–12 (2017)
7. Salmon, P.M.: Distributed situation awareness: advances in theory, measurement and application to team work. Dissertation, Brunel University, London (2008)
8. Liu, S., Wanyan, X.R., Zhuang, D.M.: Modeling the situation awareness by the analysis of cognitive process. *Bio-Med. Mater. Eng.* **24**(6), 2311–2318 (2014)
9. Smith, K., Hancock, P.A.: Situation awareness is adaptive, externally directed consciousness. *Hum. Fact.: J. Hum. Fact. Ergon. Soc.* **37**(1), 137–148 (1995)
10. Bedny, G., Meister, D.: Theory of activity and situation awareness. *Int. J. Cogn. Ergon.* **3**(1), 63–72 (1999)
11. Wickens, C.D., Mccarley, J.S., Alexander, A.L., Thomas, L.C., Ambinder, M., Zheng, S.: Attention-situation awareness (A-SA) model of pilot error. In: Foyle, D.C., Hooey, B.L. (eds.) *Hum Performance Modeling in Aviation*, pp. 213–242. Lawrence Erlbaum, Mahwah (2008)
12. Kirlik, A., Strauss, R.: Situation awareness as judgment I: statistical modeling and quantitative measurement. *Int. J. Ind. Ergon.* **36**(5), 463–474 (2006)
13. Entin, E.B., Serfaty, D., Entin, E.E.: Modeling and measuring situation awareness for target-identification performance. In: Garland, D.J., Endsley, M.R. (eds.) *Experimental Analysis and Measurement of Situation Awareness*, pp. 233–242. Embry-Riddle Aeronautical University Press, Daytona Beach (1995)
14. Hooey, B.L., Gore, B.F., Wickens, C.D., Scott-Nash, S., Salud, E., Foyle, D.C.: Modeling pilot situation awareness. In: Cacciabue, P.C., Hjälm Dahl, M., Luedtke, A., Riccioli, C. (eds.) *Human Modeling in Assisted Transportation*, pp. 207–214. Springer, Heidelberg (2010). https://doi.org/10.1007/978-88-470-1821-1_22
15. Liu, S., Wanyan, X.R., Zhuang, D.M., Lu, S.C.: Situational awareness model based on attention allocation. *J Beijing Univ. Aeronaut. Astronaut.* **40**(08), 1066–1072 (2014). (in Chinese)
16. Stanton, N.A., Chambers, P.R.G., Piggott, J.: Situational awareness and safety. *Saf. Sci.* **39**(3), 189–204 (2001)
17. Feng, C.Y., Wanyan, X.R., Chen, H., Zhuang, D.M.: Research on situation awareness model and its application based on multiple-resource load theory. *J. Beijing Univ. Aeronaut. Astronaut.* (2018). (in Chinese). <https://doi.org/10.13700/j.bh.1001-5965.2017.0532>
18. Liang, S.F.M., Rau, C.L., Tsai, P.F., Chen, W.S.: Validation of a task demand measure for predicting mental workloads of physical therapists. *Int. J. Ind. Ergon.* **44**(5), 747–752 (2014). <https://doi.org/10.1016/j.ergon.2014.08.002>



The Effect of Thirty-Six Hour Total Sleep Deprivation on Spatial Cognition and Alertness

Wenjuan Feng, Ruishan Sun^(✉), and Kai Zhang

Research Institute of Civil Aviation Safety, Civil Aviation University of China,
Tianjin 300300, China

wjfeng1121@hotmail.com, sunrsh@hotmail.com

Abstract. Objective: To explore the effect of total sleep deprivation (TSD) for 36 h on spatial cognitive ability and alertness in normal youth. **Methods:** Six healthy young men aged 22–26 were enrolled in this study. Mental rotation tests and KSS measurements were performed once every hour under 36 h TSD. **Results:** Accuracy of the mental rotation ability test first increases and then decreases during the day. KSS scores increase with TSD time in a 12-h cycle. **Conclusion:** Under the condition of total sleep deprivation for 36 h, the spatial cognitive ability of normal youth declines to a certain extent, and their own learning and proficiency greatly influences mental rotation test scores. Alertness of normal youth continued to decline over 36 h of TSD, and fatigue gradually increased.

Keywords: TSD · Spatial cognitive ability · Alertness · Biological rhythm
Aviation safety

1 Introduction

In recent years, the Chinese civil aviation industry has been greatly developed, and the total transport turnover is second only to the United States [1]. Civil Aviation Administration of China statistics shows: In 2016, the national airport of civil aviation transport completed a passenger throughput of 1.016 billion passenger trips, an increase of 11.1% over the previous year [2]. The number of civil aircraft airports completed 9,238,000 movements, an increase of 7.9% over the previous year. The performance of civil aviation has increased year by year, which has led to the continuous operation of civil aviation workers under high load conditions or continuous work around the clock, which seriously affects the safe operation of civil aviation. Considering the example of civil aviation air traffic controllers, first-line air traffic controllers adopt a 24-h shift system. Long-term shifts, heavy work intensity, time pressure, and work environment can easily result in sleep deprivation and fatigue, which jeopardizes control, work safety, and efficiency [3].

To perform efficiently, the controller must command the height, speed, and direction of the plane in three-dimensional space, and the pilot must operate the aircraft in three-dimensional space. It is necessary to have good spatial cognitive ability [4]. Visual image is an important aspect of spatial cognitive ability and may be manifested in many different ways, the most typical being appearance-based mental rotation [5]. Mental rotation is not only a typical cognitive activity of visual space but also one of the main indicators to assess the cognitive level of directional orientation in flight space [6]. During total sleep deprivation (TSD) there is no sleep for at least 24 h. This leads to a series of changes in emotion, learning and memory, immune function, etc. With increase in fatigue, a series of physiological, psychological and even behavioral changes is manifested [7]. Therefore, it is particularly important to study the changes of mental rotation and alertness of normal youth under TSD. The present study aimed to explore the effects of 36 h TSD on mental rotation and alertness of normal young individuals.

2 Method

2.1 Participants

Six healthy male youth, aged 22–26 years; height 178 ± 3.7 cm, weight 70 ± 5 kg; physical health; no sleep disorders, mental illnesses, or family history of these diseases; regular sleep routine (sleep at about 22:30 and wake up at 7:30); no recent medications, were included in the study. All participants agreed and signed informed consent before participating in this experiment. And this research was approved by Civil Aviation University of China.

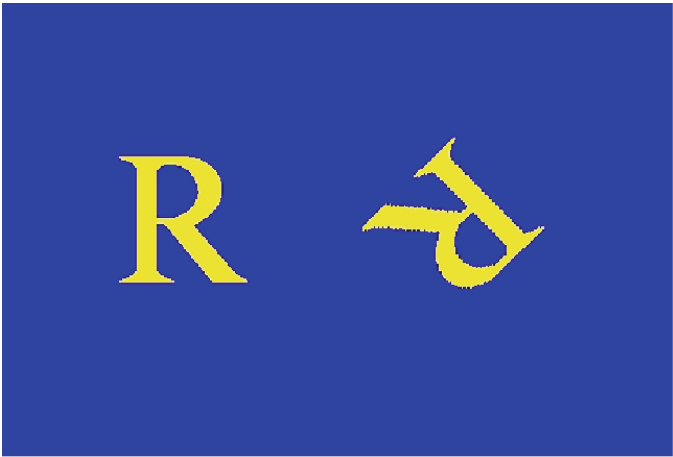
2.2 Procedures

For one week prior to the experiment, all experimental participants were directed to rest on schedule. They were banned from consuming alcohol, strong tea or coffee, performing strenuous exercise, and were prohibited from taking drugs that inhibit or excite the central nervous system. In the two days prior to the experiment, experimental procedures were explained and simulated to allow the participants to reach a degree of familiarity with the experimental procedures and instruments, and to achieve better experimental results. After two days of normal sleep, experimental procedures began at 8:00 on the day of the experiment, and ended the following day at 20:00. Mental rotation tests and Karolinska sleepiness scale (KSS) measurements were administered every hour, and total sleep deprivation (TSD) for 36 h was achieved. Six staff members took turns to supervise the participants and prevented them from taking a nap (TSD quality is considered substandard if napping occurs for more than 3 min).

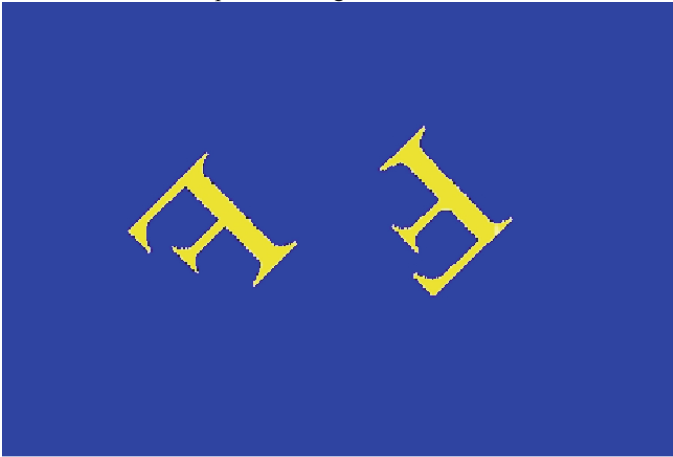
2.3 Apparatus

Test of mental rotation ability (TMRA) was performed using a DXC-6 wireless multi-group psychological evaluation instrument, made by the Department of Aerospace

Medicine of the Fourth Military Medical University of China. The test was expressed by the English capital letters G, F, and R. The two letters were about $1.0^{\circ} \times 7.9^{\circ}$, and the shape area was not more than $1.58 \text{ cm} \times 6.66 \text{ cm}$. The screen presents a pair of the same English letters, two “G”, two “F”, or two “R”, which had either a positive image or mirror image relationship. The participant was asked whether the rotation of one letter would allow it to perfectly superimpose the other (as was the case for the positive image). If this was the case, they were asked to press the “yes” key; otherwise, if the letters were mirror images, they were asked to press the “no” key.



A positive image of the letters



A mirror image of the letters

TMRA is stimulated by visual channels to respond to motion, to detect spatial cognitive ability, psychological appearance, and judgment and reasoning ability. During the experiment, six participants were tested at the same time and measurements were

performed every hour. The participants were asked to judge and answer questions in the shortest time possible. The computer automatically recorded the time elapsed from the presentation of the stimulus to the participant's response and the correct number of responses. The difficulty levels of the questions were approximately the same, and the correct number of answers was selected as the experimental indicator to characterize the participants' spatial cognitive ability.

There are five options for Karolinska sleepiness scale (KSS): 1-very alert; 2-alert; 3-general; 4-sleepy; 5-very sleepy. Participants rated their sleepiness on the scale of 1–5.

2.4 Statistical Analysis

Six participants successfully completed the experiment, and none of the participants were permitted to exit during the experiment. TMRA data includes the correct number of answers, and KSS data includes the participants' self-assessment drowsiness values. Data analysis was carried out using SPSS 11.0 statistical package, MATLAB R2014a for data processing, measurement data was represented by $(x \pm s)$, $p < 0.05$ indicated statistical significance.

3 Result

3.1 Effect of 36 h TSD on Mental Rotation Ability

Under the conditions of 36 h TSD, there are obvious individual differences in the scores of TMRA, as well as considerable differences in individual spatial cognitive ability, as shown in Table 1.

Table 1. The average number of correct answers for 6 participants

| | No. 1 participant | No. 2 participant | No. 3 participant | No. 4 participant | No. 5 participant | No. 6 participant |
|-----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| The correct number | 50.47 ± 4.99 | 62.31 ± 2.62 | 67.64 ± 2.42 | 59.5 ± 3.92 | 58.44 ± 3.62 | 62.03 ± 2.86 |

By comparing the individuals' mental rotation scores, we find that during the 36 h TSD, the mental rotation scores of participants 2, 4, 5, and 6 have roughly the same trend over time, and the individual mental rotation test scores are analyzed as follows (Fig. 1):

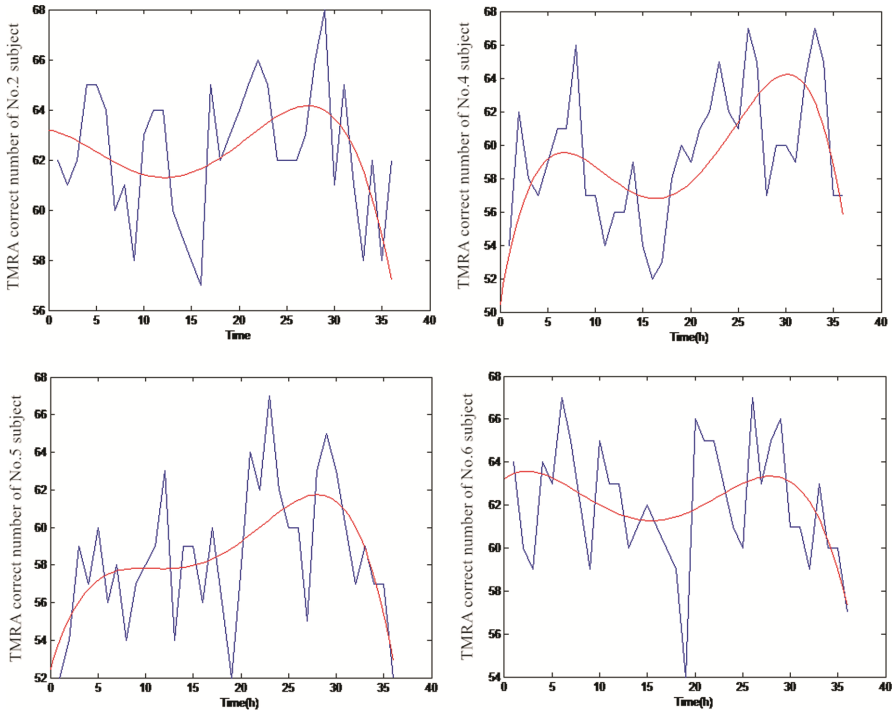


Fig. 1. TMRA scores distribution and fitting curve

During the first day (08:00 to 24:00) and the next day (00:00 to 19:00), the individuals' mental rotation performance generally showed an upward trend first and then a downward trend. At 23:00 on the same day and at 2:00 the next day the results reached a minimum.

The average answer accuracy rate of four participants during the 36 h TSD is shown in Fig. 2. In order to eliminate the effects of biological rhythms and other factors and keep the rhythm consistent, Take two days at the same time, four participants were compared the correct rate of mental rotation test. Four time points—9:00, 12:00, 15:00, and 18:00—were selected for comparison, as shown in Fig. 3.

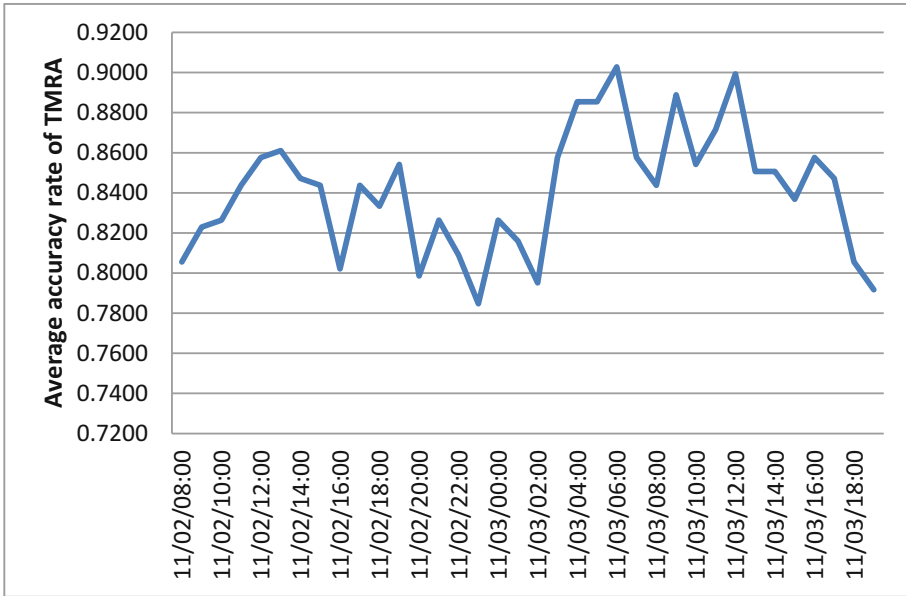


Fig. 2. Effect of 36 h TSD on TMRA average accuracy rate of 4 participants

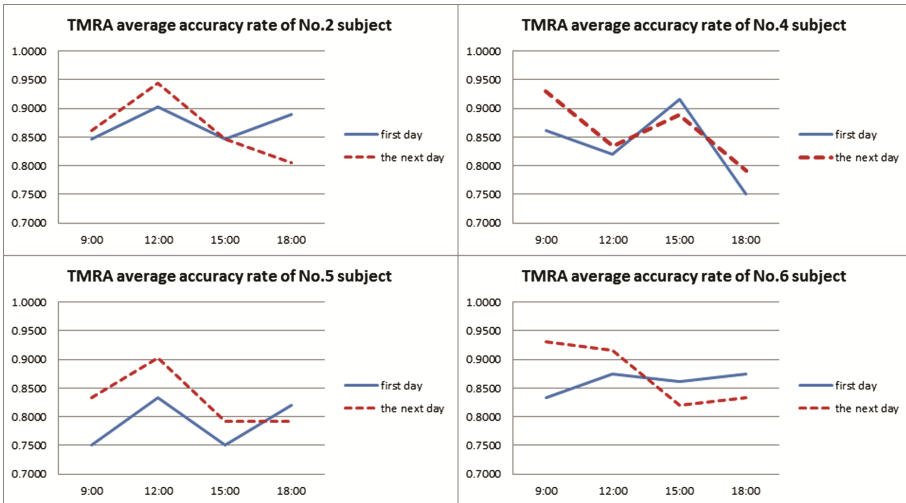


Fig. 3. Average accuracy rate of TMRA contrast

As shown in Fig. 2, the results of the mental rotation test show a pattern of first rising and then falling in a day’s time, with performance reaching a minimum at 23:00 on the first day, and 2:00 the following day. As shown in Fig. 3, comparison showed that, excluding individual biological rhythms and other factors, the mental rotation test results of four participants are higher on the second day of sleep deprivation than those during

the first day from 8:00 to 12:00. On the whole, during the 9:00 to 18:00 time period, the second day of sleep deprivation does not significantly decrease mental rotational test scores as compared with most of the corresponding time points on the first day. The results suggest that because of the large number of repetitions and simplicity of the test questions, the participants themselves became more and more familiar with the test questions. Consequently, there is no obvious declining trend during daytime hours, or on the second day compared with the first day.

3.2 Effect of 36 h TSD on Alertness

With the increase of 36 h TSD, the overall KSS score shows a fluctuating trend that is rising overall. The KSS score increases significantly ($p < 0.01$) at 13 :00, 14:00, 15:00, and 16:00 on the first day of the experiment, and 1:00, 2:00, 3:00, and 4:00, as well as at 13:00, 14:00, 15:00, and 16:00 on the second day, as shown in Fig. 4.

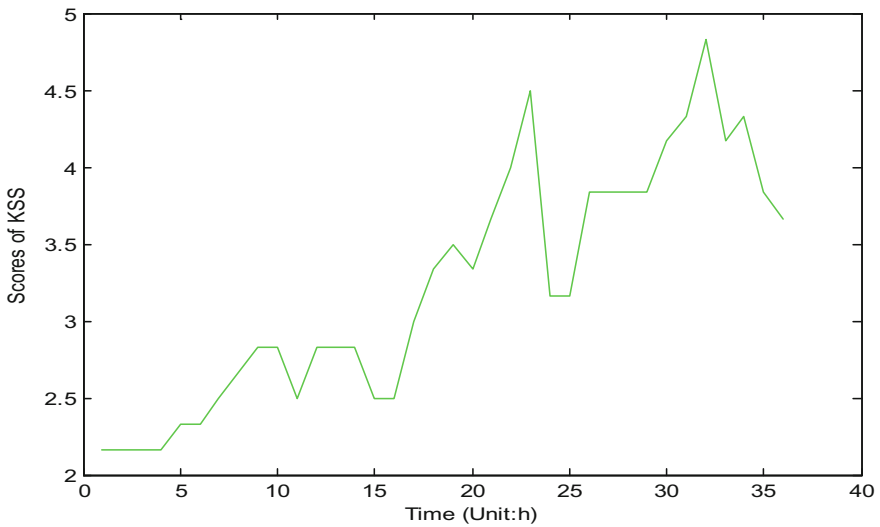


Fig. 4. Effect of 36 h TSD on the KSS scores

In this experiment, The KSS self-rating scores significantly correlated with the correct rate of mental rotation tests ($p < 0.05$). The KSS scores increase significantly from 23:00 on the first day, and begin to decrease from 6:00 the next day. The scores of the mental rotation test are at their lowest from 23:00 to 2:00 on the first day. Sleep deprivation can lead to a rising trend of alertness among normal youth. During the day, alertness scores had somewhat risen. These results suggest that the effects of circadian rhythms will offset some of the effects of sleep deprivation.

| Correlations | | KSS | Mental rotation test |
|----------------------|---------------------|-------|----------------------|
| KSS scores | Pearson correlation | 1 | .417* |
| | Sig. (2-tailed) | | .012 |
| | N | 36 | 36 |
| Mental rotation Test | Pearson correlation | .417* | 1 |
| | Sig. (2-tailed) | .012 | |
| | N | 36 | 36 |

*Correlation is significant at the 0.05 level (2-tailed).

Periodic increases in alertness are observed during TSD because of the effect of circadian rhythms. Circadian rhythm refers to the psychology and physiological functions of the human body that change over an approximately 24 h cycle. In the psychological function of the circadian rhythmic trough, alertness, perceptual ability, sustained attention ability, attention distribution and transfer ability, memory, speed of reaction, and thinking ability decline, while sleepiness increases and emotional aggravation is observed. Physiological rhythm causes the body’s hormone secretion, body temperature and other physiological processes to show periodic changes with the circadian rhythm, resulting in the fluctuation of spatial operation and cognitive ability level, as shown in Fig. 5.

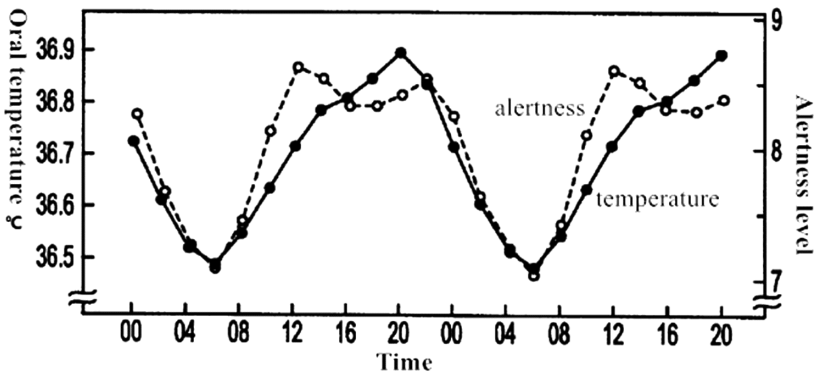


Fig. 5. The alertness and temperature vary with the rhythm.

4 Discussion

According to previous studies [8–10], TSD has less effect on physical ability and simple mental rotation ability, but has significant effect on complex mental ability. In relatively low-load tasks, physical-based, simple tasks, and tasks for which the individual is proficient are rarely affected by TSD, and more complex cognitive tasks are more affected by TSD. Cognitive tasks during simple visual search experiments show that the speed of completion of the task was significantly affected by the accuracy of little effect,

showing rhythmic fluctuations over 48 h TSD [11]. Tests on visual scanning arousal state, selective visual reaction time, memory matching operation task, exercise learning and memory ability and emotional state showed that 72 h TSD and environmental stress significantly impaired the cognitive operation ability and emotional state [12]. Thomas's study has shown that lack of sleep for a week leads to a decrease in the speed of the psychomotor alert task test and an increase in errors [13].

In the present study, spatial cognitive ability is the essential psychological quality of most aviation workers. It is a central element to ensure completing the task safely and efficiency. And the 36-h TSD can affect cognitive ability and alertness to rhythmic fluctuations. However, research shows that TSD has less impact on simple mental work, the task of mental rotation is simpler than that of complicated mental work, and the repetition rate is high. Under the conditions of 36 h TSD, the damage to spatial cognitive ability does not change significantly, and the influence of repeated operation is not ruled out. In future studies, the complexity of mental tasks may be increased, the repeatability of experiments may be avoided, the number of participants may be enlarged, and the influence of TSD on cognitive ability, judgment ability, and operation ability may be further studied.

5 Conclusion

Conclusions are summarized as follows:

There are obvious individual differences in the spatial cognition of normal young people in a certain degree. Under the condition of 36 h TSD, the spatial cognitive ability of normal young people is decreased to a certain degree. Performance is the lowest between 23:00 on the first day and 2:00 on the second day. Because of the test simplicity and the number of repetitions, the learning proficiency of the test subjects has a certain impact on the mental rotation performance. As such, their scores on the mental rotation test during the daytime hours on the second day did not drop significantly as compared with those obtained on the first day. Thirty-six hour TSD causes a rhythmic fluctuation decline in the alertness of normal youth, however during the daytime hours the effects of sleep deprivation are offset by a rise in biological rhythms, restoring alertness.

Lack of sleep and circadian rhythms of the human body are important factors that affect aviation safety, will lead to alertness, perceptual ability, sustained attention ability, attention distribution and transfer ability, memory and speed of reaction decline, while sleepiness increases and emotional aggravation is observed. Therefore, to counter the adverse effects of TSD, corresponding measures should be taken when the time of TSD is long and occurs during the trough of individual biological rhythms. For flight crew members, sufficient rest and sleep should be performed before across time zones and long-range flight. And should be scientifically scheduled and arranged with a sufficient number of pilots. Crew members take turns napping for 40 min, helping to relieve fatigue and maintain endurance during the rest of the flight. After arriving in the destination, they should rest for a sufficient time to complete the rhythm reconstruction. For air traffic controllers, if they lack enough sleep, the possibility of fatigue will rise sharply, even asleep during work. Efficient schedules will rationalize resources and mobilize

resources, so that controllers can have better rest. All the aviation staff should be scheduling to prevent fatigue, and put the fatigue risk management into practice.

References

1. Wang, X.: International experience on deregulation of air transport industry and its enlightenment to China. *Civ. Aviat. Manag.* (11), 31–33 (2011). (Chinese)
2. Civil Aviation Administration of China: Statistical Communique on the Development of Civil Aviation in 2016 (2017). (Chinese)
3. Transport Canada: Fatigue in air traffic controllers: literature review. *Fatigue* (2000)
4. Sladky, R., Stepniczka, I., Boland, E., Tik, M., Lamm, C., Hoffmann, A., et al.: Neurobiological differences in mental rotation and instrument interpretation in airline pilots. *Sci Rep.* **6**, 28104 (2016)
5. Qi, L.: Image representation: researches of mental rotation. Doctoral dissertation, East China Normal University (2009). (Chinese)
6. Wang, P., Huang, Y.: Survey of mental rotation. *J. NINGBO Univ. (Educ. Sci.)* **28**(5), 20–22 (2006)
7. Killgore, W., Kahn-Greene, E.N., Killgore, D., Balkin, T.: Sustaining executive functions during sleep deprivation: a comparison of caffeine, dextroamphetamine, and modafinil. *Sleep* **32**(2), 205–216 (2009)
8. Heuer, H., Kleinsorge, T., Klein, W., Kohlsch, O.: Total sleep deprivation increases the costs of shifting between simple cognitive tasks. *Acta Physiol.* **117**(1), 29–64 (2004)
9. Hood, B., Bruck, D.: A comparison of sleep deprivation and narcolepsy in terms of complex cognitive performance and subjective sleepiness. *Sleep Med.* **3**(3), 259 (2002)
10. Howard, S.K., Gaba, D.M., Smith, B.E., Weinger, M.B., Herndon, C., Keshavacharya, S., et al.: Simulation study of rested versus sleep-deprived anesthesiologists. *Anesthesiology* **98**(6), 1345 (2003)
11. Li, Y.F., Zhan, H., Li, T.: Effects of 48 h sleep deprivation on dual task ability and fatigue. *Chin. J. Appl. Physiol.* **21**(2), 174–5, 191 (2005). (Chinese)
12. Lieberman, H.R., Tharion, W.J., Shukitt-Hale, B., Speckman, K.L., Tulley, R.: Effects of caffeine, sleep loss, and stress on cognitive performance and mood during U.S. Navy SEAL training. *Psychopharmacology* **164**(3), 250–261 (2002)
13. Thomas, M., Sing, H., Belenky, G., Holcomb, H., Mayberg, H., Dannals, R., et al.: Neural basis of alertness and cognitive performance impairments during sleepiness ii. effects of 48 and 72 h of sleep deprivation on waking human regional brain activity. *Thalamus Relat. Syst.* **2**(3), 199–229 (2003)



Human-Centered Design of Flight Mode Annunciation for Instantaneous Mode Awareness

Andreas Horn, Wen-Chin Li^(✉), and Graham Braithwaite

Safety and Accident Investigation Centre, Transportation Division,
Cranfield University, Bedford, UK
wenchin.li@cranfield.ac.uk

Abstract. Since the early days of aviation, aircraft design engineers have been using automation to reduce pilot workload and enhance flight safety. While the basic automated systems performed quite simple tasks such as to hold altitude or heading, modern flight guidance and control systems typically have over 20 different modes of operation. The basic philosophy of flight mode annunciation however, did hardly change. It is therefore not surprising that flight crews encounter more and more difficulties to comprehend the active flight mode. In current designs, the active flight mode is typically shown on top of the primary flight display (PFD). A new flight mode annunciation (FMA) concept was investigated using a flight simulator in conjunction with eye-tracking analysis and a questionnaire. The experiment involved 20 participants, aged between 22 and 47 years ($M = 27.7$, $SD = 7$). Thereof 12 were qualified pilots and 8 were aerospace engineering professionals. Flight experience reached from 0 to 11000 h ($M = 946$, $SD = 2567$). The study showed that the modified display style caused a significant decrease of fixation duration and subjective workload.

Keywords: Flight mode annunciation · Mode confusion
Human-computer interaction · Automation · Situation awareness

1 Introduction

The continuous reoccurring of accidents and incidents involving insufficient situational awareness and mode confusion underlines the need in the aerospace industry to develop a more simplistic and easy to interpret way of flight mode annunciation. A well-known example of such an event is the accident involving a Boeing 777 aircraft at SFO airport. During a visual approach in clear weather conditions, flight crew actions led to several mode changes of the automation, that were not perceived and interpreted correctly by the flight crew, ultimately resulting in the aircraft hitting the sea wall short of the runway [1]. The investigation clearly showed, that the flight crew did not have sufficient awareness of the current status of the automation [1]. The proposed concept was implemented on a Boeing 777 PFD, in order to investigate how it would affect the effort of interpretation in similar situations.

The key idea behind the modified display style is to merge the FMA with raw flight data on the PFD and thus to embed it in the natural scanning pattern of a pilot.

Additionally, the cognitive work of interpreting the FMA and correlate it with the raw data shall be significantly reduced by simply showing with a “green border”, whether or not a certain parameter is controlled by the autopilot flight director system (AFDS). Björklund et al. [2] found that pilots are paying very little attention to the FMA when *not expecting* automation mode changes. The accident in SFO confirmed, that a flight crew can easily miss such an unexpected change [1]. A conventional PFD (B777) as shown in Fig. 1 displays raw flight data, such as airspeed (left), attitude (center), altitude (right) and navigation information (i.e. ILS deviation scales). The FMA is located on top of the display and shows “active” modes in green, as well as “armed” modes in white. The FMA is divided into three columns for auto throttle, roll-mode and pitch-mode and one AFDS status field [1]. The red boxes in Fig. 1 list all the possible modes for each column and the AFDS status field respectively. The high cognitive effort required for interpretation can be demonstrated as follows: Assuming the flight crew desires to find out, if the airspeed is automatically controlled, they need to do the following:

- Read the FMA auto throttle column text
- Interpret the text, as there are modes that cause the auto throttle to be “engaged”, but not controlling the airspeed (e.g. “HOLD” or “THR”)
- If the text is anything else than “SPD”, the scanning continues to the pitch-mode, as this channel can also be used to control the airspeed [1]
- Interpret the pitch-mode (check if it is FLCH SPD or VNAV SPD)
- Check the AFDS status indication, to ensure that the autopilot is engaged in case of airspeed being controlled via the pitch-mode.



Fig. 1. Complex nature of flight mode annunciation in current display style (Color figure online)

The example of Fig. 1 demonstrates that the complex nature of control-mode interdependence of current PFD creates a situation that cannot be readily conveyed to pilots using text only. One of the major problems with the current flight mode annunciation philosophy is that the basic layout did not change for decades, while the capabilities of the automation systems evolved dramatically [3]. For a basic function, like “altitude hold” or “heading hold” a very basic annunciation concept using text labels was capable of providing adequate information. With the advent of more sophisticated automation concepts and the introduction of the auto throttle, the situation became much more complex. A basic flight parameter such as indicated airspeed can now be controlled by multiple systems (i.e. autopilot pitch mode or auto throttle) [3]. The location of the FMA on top of the PFD is in peripheral vision most of the time. The current FMA however, is designed for foveal vision in terms of the saliency it provides [4]. Furthermore, the current philosophy does not allow the flight crew to readily identify if a certain parameter is controlled by the automation [3], as the cognitive task of relating the flight mode annunciation text to the physical behaviour of the aircraft is left to the flight crew. This has led to several occurrences of so-called “controlled flight into stall” (CFIS) [3]. Sherry and Feary [5] found that another cause for inaccurate understanding of automation behaviour lies in the differences between flight crew training documents and manufacturers technical specifications. These differences are then often rectified by gaining practical experience during the line-training [6].

The proposed modified design shown as Fig. 2 aims to remedy this by relating the flight mode annunciation to the physical flight parameters. This new PFD layout consists of only small changes to the general appearance in order to keep transition training for pilots to a minimum. At the same time there is a distinct change in philosophy: Instead of just displaying the flight mode as in the legacy design, the new design also highlights the parameter that is controlled. In simple terms, this could be described as showing *what* is controlled, rather than just *how* it is controlled. The basic FMA box is retained, while green borders are added for the parameters that are actively controlled by the automation. In Fig. 2 the active flight modes “SPD”, “LOC” and “G/S” are augmented with the respective parameters being highlighted. It is therefore instantly clear, whether or not a certain parameter is controlled by the automation, simply by looking for the presence or absence of green borders.



Fig. 2. Proposed modified flight mode annunciator by green color border (Color figure online)

2 Method

2.1 Participants

The experiment involved 20 participants, aged between 22 and 47 years ($M = 27.7$, $SD = 7$). Thereof 12 were qualified pilots and 8 were aerospace engineering professionals. Flight experience reached from 0 to 11000 h ($M = 946$, $SD = 2567$). All participants were provided with a consent form before the experiment. Furthermore, the experiment process was reviewed and approved by the Cranfield University Research Ethics System (CURES), reference number 2475.

2.2 Apparatus

Flight Simulator: A virtual replica of the B777 instrument panel was used to create the basic scenarios. The Precision Manuals Development Group (PMDG) “B777 expansion pack” allowed authentic recreation of the B777 PFD and ND. The experiment was conducted in a segregated room that was quiet and free from optical disturbances. Windows were blanked off and the illumination level kept constant.

Eye Tracker “Pupil” is a wearable, light-weight eye-tracking device that can be used for automated eye-movement analysis in everyday life [7]. It consists of a headset including cameras and software packages for capture and analysis. The headset is connected to any convenient computing device (e.g. laptop) using an USB connection. The Laser-sintered headset hosts two cameras, one facing the right eye of the participant (eye-camera), the other capturing the field of vision (scene-camera) [7]. The eye-camera has a resolution of 800×600 pixel and a frame rate of 60 Hz. The detection of the pupil is based on the “dark-pupil” concept, using the infrared spectrum. The scene-camera captures the user’s field-of-view at a high-resolution (1920×1080 pixel) with a frame rate of 60 Hz.

2.3 Scenario

The research involved developing a new display concept for flight mode annunciation and verifying it using an eye-tracking experiment and a questionnaire. Five typical scenarios were developed, using both the current and modified display style for each of them. Table 1 depicts the essential description of each simulated stage of flight.

Table 1. Simulated stages of flight

| Scenario | Task |
|----------|--|
| 1 | ILS approach |
| 2 | Climbing turn |
| 3 | Descending turn |
| 4 | Level turn with subsequent descent |
| 5 | Desired track intercept (level flight) |

2.4 Hypothesis

Of particular interest is the flight crew workload during the task and the ease of interpretation of the FMA. The combination of objective (eye-tracking) and subjective data (NASA-TLX) serves as a basis for this assessment. These two designs (current and modified) are compared using the following set of experimental null hypotheses:

1. There is no significant difference in fixation duration
2. There is no significant difference in pupil size
3. There is no significant difference in perceived workload.

2.5 Research Design

Participants were split into qualified and unqualified groups in order to allow a mixed-design analysis to be carried out. The qualified group included all participants with any kind of flight experience, reaching from some flight training on single engine piston aircraft up to senior captain of multi-engine jet aircraft. The unqualified group consisted of all other participants, mainly aerospace engineering professionals. In order to counterbalance the “learning effect” each participant may experience during the experiment and also to remedy any bias based on the sequence of display styles, the scenarios and display styles were randomly assigned to participants. Following the guidance provided in [8], a “Williams design” was developed for the five scenarios and two display styles. Two dedicated pilot tasks were created to generate a realistic workload. The key aspect here is that the participants were told to be the “pilot monitoring”, simulating a multi-crew environment and requiring them to check the progress of the flight with respect to given constraints. This represents a typical real-world situation where the aircraft is controlled by the “pilot flying” or the autopilot and the “pilot monitoring” has to verify the adherence to published procedures. Additionally this was deemed to create an acceptable amount of cognitive workload apart from monitoring the FMA, in order to avoid prolonged fixation on the FMA.

The first task involved monitoring of airspeed. Subjects were asked to callout every 10 kts of change in airspeed. As a second task, any altitude change of 100 ft had to be called out. These two tasks existed in addition to monitoring of the FMA. Any change on the FMA had to be called out and was recorded by the operator. As not all of the participants were familiar with the intricacies of the B777 flight modes, the emphasis was laid on the notification of the flight mode text change, rather than the understanding of the physical meaning of the respective flight mode. All participants evaluated their perceived workload using NASA-TLX after each scenario.

3 Results and Discussions

A mixed-design repeated measures two-way ANOVA was carried out for the eye-tracking parameters listed in this section. The within-subject factors were display style of FMA (current and modified) and scenario (1–5), while flight experience (experienced vs non-experienced) was used as a between-subject factor.

3.1 Fixation Duration

There is a significant difference on fixation duration between conventional FMA and modified FMA, $F(1, 16) = 6.81$, $p < .01$, partial $\eta^2 = 0.299$. Evidence of a significantly lower fixation duration was found for the modified display style. The greater amount of saliency provided by the modified display style seems to facilitate the process of comprehending a new automation state, possibly due to the related change in the participant's saliency map [9]. The lower fixation duration also serves as an objective confirmation of the faster processing time and lower subjective workload ratings mentioned by many participants. It was observed that the qualified group had a smaller variation in fixation duration than the unqualified group. One of the key elements in pilot training is to establish a useful scanning pattern and avoiding to fixate too long on only one specific area of the PFD [10]. It is very likely that this training bias manifests itself in the smaller and more consistent fixation duration. Differences in fixation duration between experienced and novice pilots were also observed by Yu et al. [11]. The null hypothesis for fixation duration can therefore be rejected.

3.2 Pupil Size

There is no significant difference on pupil dilation between conventional FMA and modified FMA, $F(1, 16) = 2.67$, $p = 0.125$, partial $\eta^2 = 0.141$. Although a smaller pupil size was measured in modified PFD, this effect was not significant. Ahlstrom and Friedman-Berg [12] found that pupil-size *does* change with workload, however the change in pupil-size for the medium-workload region was minimal. It was therefore not possible to underline the lower subjective workload with this measurement. The unqualified group had a much smaller pupil size. It should be noted that the parameter of interest is the *change* in pupil size rather than the absolute value of the pupil size [13]. Yu et al. [11] found significant differences in pupil size of pilots depending on the flight scenario and task. The observed effect shows that it is possible to “merge” the flight mode annunciation with the raw flight data and thus co-locate important information. This facilitates the scanning pattern and improves the perception of information in accordance with the proximity compatibility principle [14, 15].

3.3 Subjective Workload

There is a significant difference on subjective workload between conventional FMA and modified FMA, $F(1, 16) = 11.67$, $p < .01$, partial $\eta^2 = 0.422$. Both pilot groups showed a lower subjective workload when using the modified display style. The merging of raw data with flight mode annunciation data seems to reduce the perceived effort in interpreting the displayed information. It should be mentioned, that the unqualified group adapted much better to the modified display style, than the qualified group. This is not surprising, bearing in mind that the qualified pilots have undergone significant hours of training using the legacy display philosophy. The simple design of

the modified display style allowed for a rapid transition and quick adaptation even for the experienced pilots. A very significant difference in subjective workload allows the rejection of the corresponding null hypothesis.

3.4 Scenario and Style Results Summary

The overall effects of the modified display style caused a decrease in all three cases, although statistical significance was only reached for Fixation duration and subjective workload, as described in the previous sections. The use of objective (eye-tracking) and subjective criteria (NASA-TLX) combined proved to be a useful hybrid solution, aiming to eliminate any bias in either of the parameters. The experiment also showed that the introduction of a modified display style does not show equal benefits for each one a given set of different flight situations. In particular, it could be shown that the current design works quite satisfactorily when the workload is low (scenario 5). However, if the workload gets high, the increased scanning demand causes a deterioration of mode awareness. This is precisely the situation where the modified display style is fundamentally different and advantageous, as it incorporates the automation state in the natural scanning path and thus reduces the crew effort to establish a good mode awareness.

3.5 Attention Distribution

Based on the eye-tracking data, so-called “heat-maps” were calculated, depicting the gaze distribution (Red reflects large amount of time, relative to the total scenario duration). Figure 3 represents a classic example of a professional pilot’s behaviour. Attention is given to the raw data fields primarily, and the flight modes were only checked occasionally. It is worth noting that the particular subject was able to maintain a good situational awareness throughout all scenarios. Most of the saccades performed towards the FMA seem to be based on *expectancy* of a mode change. This shows the vulnerability to miss “uncommanded” or automation-induced mode changes. The modified FMA is much more likely to get the attention of the pilot, as the green borders appear/disappear directly around the raw data fields shown as Fig. 3. It cannot be over-emphasized that scanning the raw data is important and crucial for the safety of flight. It really is the limited scanning of the FMA box that should raise concerns about the conventional design.

Figure 4 depicts a typical sequence of fixations during the experiment. After checking the AFDS status field (1) the subject scans the altitude (2) and the airspeed (3), before reaching the LNAV deviation scale (4). This example also underlines the need for the automation mode annunciation to be incorporated in the raw data field, as it is obvious from the picture, that scanning raw data necessitates a diversion from the FMA in the current design philosophy. In fact it can be seen, that by fixation (3) and (4), the pilot automatically gets the current automation state with the new design using the green borders.

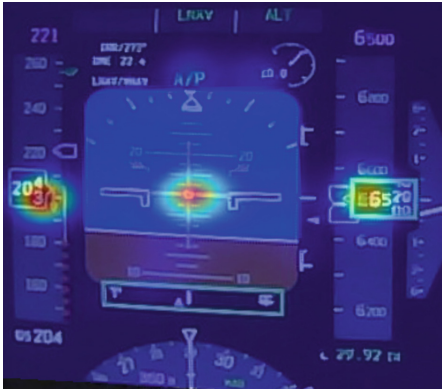


Fig. 3. Heatmap of attention distribution. (Color figure online)

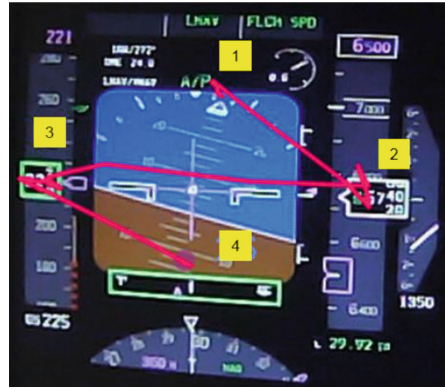


Fig. 4. Scanning path for 3 s (Color figure online)

4 Conclusion

The experiment showed that the modified display style causes a significant reduction in workload, by applying the proximity compatibility principle [14]. The flight crew's ability to rapidly gain situational awareness is vital for flight safety. The modified display style reduced the cognitive and temporal workload when interpreting the automation status on the PFD. Additionally, the modified display style is much more likely to catch the flight crew's attention when an "unexpected" mode change occurs, as the salient stimulus is applied directly in the raw data fields, which are frequently looked at by the flight crew. The proposed method was tested for the "autopilot engaged" situations only during this experiment, however an expansion to the "flight director only" regimes is also possible. It should also be noted that the use of green borders is by no means a very sophisticated way of incorporating the automation state in the raw data fields. The use of color-coded or solid and hollow deviation bars and indications would be a next design step. This was also underlined by subjective qualitative feedback obtained from participants. Professional pilots highly appreciated the idea of the concept, but provided critical comments on the implementation. An improved experiment design could incorporate a "trial" phase first, to eliminate most of the implementation problems before the actual experiment.

An interesting discussion developed when participants were asked about an aural flight mode annunciation: The professional pilots were strongly opposing this idea. Nevertheless research in this area should continue, as the introduction of data link communication for air traffic control [16] will only further overload the optical channel and reduce the use of aural information exchange between the pilots and air traffic control (ATC). Furthermore, Zhang et al. [17] demonstrated that a timely aural warning can significantly improve a human's ability to quickly fixate on a critical object. A good example of combining aural advisories with visual perception in a cockpit has been tested by Purcell and Andre [18]. They were able to reduce head-down-time during taxi operations by using an aural cue when to look at the moving map. A similar

concept could also work for the FMA. Caution should be exercised when using similar aural alerts for multiple different events. Kearney et al. [19] found that in the case of air traffic controllers, the acoustic alert led to a slower reaction than a semantic alert, as the same aural sound was used in multiple instances.

References

1. NTSB: Descent Below Visual Glidepath and Impact with Seawall, Asiana Airlines Flight 214, Boeing 777-200ER, HL7742, San Francisco, California, 6 July 2013. NTSB (2014)
2. Björklund, C.M., Alfredson, J., Dekker, S.W.A.: Mode monitoring and call-outs: an eye-tracking study of two-crew automated flight deck operations. *Int. J. Aviat. Psychol.* **16** (3), 263–275 (2006)
3. Sherry, L., Mauro, R.: Design of Cockpit displays to explicitly support flight crew intervention tasks. In: 2014 IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC), pp. 2B5–1 (2014)
4. Nikolic, M.I., Sarter, N.B.: Peripheral visual feedback: a powerful means of supporting effective attention allocation in event-driven, data-rich environments. *Hum. Factors* **43**(1), 30–38 (2001)
5. Sherry, L., Feary, M.: Evaluating flight crew operator manual documentation. In: 1998 IEEE International Conference on Systems, Man, and Cybernetics, vol. 1, pp. 897–902 (1998)
6. Holder, B., Hutchins, E.: What pilots learn about autoflight while flying on the line. In: Proceedings of the 11th International Symposium on Aviation Psychology, pp. 526–531 (2001)
7. Kassner, M., Patera, W., Bulling, A.: Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pp. 1151–1160 (2014)
8. Gong, L.-K., Wang, X.-J., Wang, B.-S.: The construction of a Williams design and randomization in cross-over clinical trials using SAS. *J. Stat. Softw.* **29**, 1–10 (2009)
9. Itti, L., Koch, C., Niebur, E., et al.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998)
10. FAA: Airplane flying handbook. FAA (2016)
11. Yu, C.-S., Wang, E.M., Li, W.-C., Braithwaite, G., Greaves, M.: Pilots' visual scan patterns and attention distribution during the pursuit of a dynamic target. *Aerosp. Med. Hum. Perform.* **87**(1), 40–47 (2016)
12. Ahlstrom, U., Friedman-Berg, F.J.: Using eye movement activity as a correlate of cognitive workload. *Int. J. Ind. Ergon.* **36**, 623–636 (2006)
13. Pomplun, M., Sunkara, S.: Pupil dilation as an indicator of cognitive workload in human-computer interaction. In: Proceedings of the International Conference on HCI, pp. 542–546 (2003)
14. Wickens, C.D., Carswell, C.M.: The proximity compatibility principle: its psychological foundation and relevance to display design. *Hum. Factors* **37**(3), 473–494 (1995)
15. Murata, A., Akazawa, T.: Basic study on automotive display design by proximity compatibility principle. In: 2014 Proceedings of the SICE Annual Conference (SICE), pp. 971–978 (2014)
16. ICAO: 2013–2028 Global Air Navigation Plan. ICAO (2013)

17. Zhang, Y., Yan, X., Li, X., Xue, Q.: Drivers' eye movements as a function of collision avoidance warning conditions in red light running scenarios. *Accid. Anal. Prev.* **96**, 185–197 (2016)
18. Purcell, K.P., Andre, A.D.: Effects of visual and audio callouts on pilot visual attention during electronic moving map use. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 44, no. 13, p. 108 (2000)
19. Kearney, P., Li, W.-C., Lin, J.J.: The impact of alerting design on air traffic controllers' response to conflict detection and resolution. *Int. J. Ind. Ergon.* **56**, 51–58 (2016)



Inter-sector Backup Behaviors in Parallel Approach ATC: The Effect of Job Satisfaction

Yazhe Li^{1,2,3}, Xiaotian E^{1,2}, Han Qiao^{1,2}, Xiangying Zou^{1,2},
Chunhui Lv^{1,2}, Lin Xiong³, Xianghong Sun^{1,2},
and Jingyu Zhang^{1,2} (✉)

¹ Institute of Psychology, Chinese Academy of Sciences, Beijing, China
zhangjingyu@psych.ac.cn

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

³ Airspace Management Center, Air Traffic Management Bureau,
Civil Aviation Administration of China, Beijing, China

Abstract. Inter-sector cooperation between air traffic controllers (ATCos) can provide an effective way to manage controllers' workload by redistributing individual task-load at a group level, yet this area has not been fully explored. Based on our previous studies which have identified the effects of certain task-level features, this present research aimed to identify the influence of individual difference variables on controllers' backup decisions. Forty licensed controllers performed thirty-two simulated final approach scenarios in which they had to decide whether to accept a hand-over request made by a controller working in the neighboring sector. We manipulated three key task-level features across scenarios: (1) participants' task-load, (2) the requestors' task-load and (3) the close-landing demands of the to-be-hand-over aircraft. We also measured controllers' work experience and job satisfaction. HLM analysis showed that: after controlling for the effects of task-level variables, job satisfaction had a unique contribution and also interacted with task-level variables in predicting backing up behaviors.

Keywords: Air traffic control · Inter-sector cooperation · Job satisfaction

1 Introduction

Ground-based air traffic controllers (ATCos) play a crucial role in guaranteeing airline efficiency and safety. This is especially the case as airspace saturation, and personnel shortage is becoming more salient [1]. Previous research has made valuable efforts to manage controller's workload from an individual-focused or with-in sector approach [2–7], but the cooperation between sectors is less investigated. While inter-sector cooperation or backup behaviors can help redistribute task load so overload at an individual level can be prevented, factors that can influence such acts have not been thoroughly investigated in the ATC domain.

Considered to be critical for effective team performance, backing up behaviors refers to helping other team members perform their roles [8]. In previous studies in the non-ATC domain, it has been found that in deciding whether to offer backup,

participants consider both their teammates' situation as well as their conditions [8–10]. Perceived workload and legitimacy are two important proximal predictors of backup behaviors. It is quite apparent that when helpers are under lower workload, they are more likely to make backup decisions. This is because the cognitive resource of the helpers is limited, they cannot offer any hand before finishing their own work. “The Legitimacy of need” was considered as the intersection of workload and capacity, and a positive relationship between the legitimacy of need and backing up behaviors was found in [8]. In a highly legitimate condition, for example, when the teammate is under high pressure, he/she will be more likely to receive backup from others.

However, to shed more light on our understanding of the backup behavior in the ATC domain, it is essential to explore how these two proximal factors are influenced by other more distal task and individual factors in the ATC domain. In the ATC domain, a typical inter-sector backup request happens when a controller (the requestor) asks his/her neighboring controller (the helper) to take-over the control of a certain amount of aircraft. The requestor should have managed these aircraft and accepting this request is beyond the helper's own duties and responsibilities. In this situation, the helper needs to decide on whether to take these aircraft or not.

There are many forms of inter-sector backup (e.g., between two neighboring en-route controllers, an en-route controller and an approach controller, and two approach controllers). Through interviewing professional expert controllers, we found that the cooperation between two parallel final approach sectors was among the most important situations in real practice. Such a configuration is common for large airports which have multiple independent runways. In a typical pair involving two parallel approach sectors, two final approach controllers manage two adjacent runways and nearby airspaces. Since the landing and take-off of aircraft in one runway can't interfere that of another, the controllers independently conduct their work. However, in certain conditions, one controller may find he/she cannot effectively manage the aircraft in his/her own sector and requests to hand over some planes in his/her sector to another one's. Facing this request, the neighboring controller needs to decide whether to accept or not. Several task properties have been further identified to influence the backup decision-making process: (1) the task load of the controller being requested, (2) the requestors' task load, and (3) the close landing demand of the to-be-hand-over aircraft.

The first factor was hypothesized to have a negative effect on one's willingness to provide backup. The reason is quite obvious as controllers need to manage their sector first. From the perspective of the Attentional Resource Theory [11, 12], if their task has highly occupied people, they cannot have additional resources to provide help to their colleagues. Indeed, lending hands to others may reduce their task performance [9]. Moreover, preventing overload is an important concern for this safety-critical occupation [5–7, 13–15]. As a result, we consider this variable to be the most important predictor of the backup decision. In experiments, this factor can be manipulated by aircraft count in the participants' sector [16–18].

The requestors' task load was hypothesized to have a positive effect on one's willingness to provide backup. This is because the controllers being requested will evaluate whether the request is reasonable. When the requestors' task load is high (e.g., many aircraft in his/her sector), his/her call for help can be perceived as legitimate. On the other hand, when the requestors' task load is low, his/her request for help can be perceived as social “loafing” or not genuine.

The close landing demand is a feature of the to-be-hand-over aircraft. When the taxiing distance of an aircraft can be reduced by being handed over to another controller's sector, we call this aircraft has a close landing demand. This happens because each airline company has its boarding gates. For example, if the boarding gates of a company are to the east, then when a plane of that company coming from the west would use the west runway for landing and taxi a quite long distance to reach its boarding gate. Of course, if such a plane can be handed over to the controller in charge of the east runway, it can land using the runway closer to its boarding gate. For controllers, this is not a requirement, but they tend to provide convenience for passengers and crews if they have enough mental capacity to deal with the overall situation. As a result, if the to-be-hand-over aircraft has a close landing demands, it is more reasonable to accept such a plane. So it is also hypothesized that controllers are more willing to accept airplanes with close-landing demands.

In a previous experiment that directly manipulated the three factors, it was found that all these task-level variables indeed influence controllers' backing up behaviors as hypothesized [19]. Specifically, the effect of controllers' task load on their backing up willingness and decisions was mediated by their perceived mental workload; the effect of the requestors' task load and the close landing demands was mediated by the perceived legitimacy of backing up requests. These findings provide initial evidence confirming the previously mentioned framework on controllers' backup behaviors. However, we also found that in addition to these task-level differences, there was a significant amount of unexplained variance that was between individual. Without a closer scrutiny of these individual differences, it is difficult to draw a firm conclusion about this behavior [2].

In the present study, we would focus on the individual difference factors that may also contribute to the decision-making process beyond the task-level factors by re-analyzing the previously collected data. Two factors will be examined in the current study. The first one is work experience. Work experience is considered as the most critical individual difference factor in ATC performance [3, 12, 20]. However, its effect on backup intentions is not self-evident. From an attentional resource perspective, controllers with more experience consume less cognitive resources in performing the same task as compared to novices. In this regard, the experienced controller might be more likely to provide backup because they have more surplus of resources to cope with other situation. However, there is also evidence that more experienced controllers also tend to make higher risk evaluation than their less experienced counterparts [3, 12]. In this way, they may think to provide help, i.e., bring additional risk to their airspace, might cause much more burden than their less experienced counterparts. Taken together, there is no substantial evidence to support a definite relationship between work experience and backup intentions. In this study, we will explore the effect of work experience on backup decisions.

The second one is job satisfaction, which is an evaluative state that expresses contentment with and positive feelings about one's job [21]. Indeed, providing inter-sector backup can be considered as a typical "citizenship" behavior, which was defined as behaviors that can "lubricate the social machinery of the organization" but beyond "the usual notion of task performance" [21]. A vast quantity of evidence collected from service industry supports a positive relationship between job satisfaction

and citizenship [21, 22]. Experiencing a generalized positive mood from their work, people are more likely to make typical citizenship behaviors such helping their colleagues and serve their customers [23, 24]. Whereas there is no direct evidence concerning the ATCos, air traffic controller with a higher level of job satisfaction might be more willing to help their colleagues and serve the passengers. So we expected job satisfaction to have a positive effect on their back up decisions.

2 Method

2.1 Participants

Forty licensed controllers from a provincial ATC center in South China participated this experiment. Their ages ranged from 22 to 48 ($M = 27.68$, $SD = 4.99$), and work experience ranged from 1 to 20 years ($M = 4.97$, $SD = 3.93$). They were paid for this and one other sub-task at a total payment of 300 Yuan.

2.2 Task and Design

Parallel Final Approach Task. We used ATC-Simulator, a medium-fidelity air-traffic-control simulation platform for research purpose [25–27] in this study. The interface was designed to simulate two parallel sectors in final approach. In this experiment, participants were asked to manage the sector on the right while the left sector was handled by another controller (whose behaviors were previously programmed). As shown in Fig. 1, the area in dark grey was participants' sector in which participants were able to manipulate the aircraft, while the light grey area was their colleague's sector which participants were not able to control.

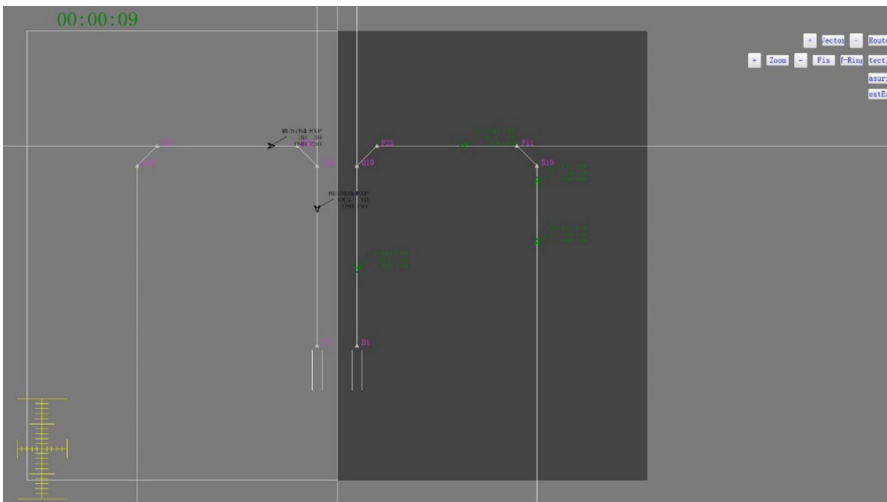


Fig. 1. The interface of the parallel approach simulator.

For each scenario, aircraft were flying into two final approach sectors. Each plane had a data block showing its call sign, flight level, bearing, and velocity. In the first 40 s, participants first needed to monitor and manage aircraft in their sector to make sure the minimum separation standard was never violated, and all the flight conditions were following the operational protocols. Then they were asked to assess their mental workload during the previous work. After that, participants were informed that the colleague managing the left sector requested to hand over one aircraft. They were asked (1) to what extent they were willing to accept this aircraft (from 1 for extremely low to 8 for extremely high, referred to as Backing-up Willingness below), and (2) if they would accept the to-be-handover aircraft or not (referred to as Backing-up Decision, Reject coded as 0 and Accept coded as 1). After completing both questions, participants were to perform the next scenario.

Aircraft were in conflict if they simultaneously violated both the lateral (5 nm) and vertical (1,000 ft) separation standard. Aircraft positions were updated every second. Two supportive tools were offered to help participants make decisions: (1) a 20 nm 10 nm scale maker in the bottom left corner which could be moved anywhere within the airspace; (2) a distance/time calculation function was provided. If participants hold down the left mouse button to select a plane and move it to any point on the screen, they would see the distance (in nautical miles) and time (in minutes) to reach that point (if maintaining the current velocity).

Scenarios. There were 38 scenarios in total, including the first six practicing scenarios and 32 formal scenarios. A 2 (*Participants' task load*: low vs. high) \times 2 (*Requestor's task load*: low vs. high) \times 2 (*Close landing demand*: no vs. yes) design was used thus creating eight different conditions. For each condition, four distinctive scenarios were created thus producing the overall 32 scenarios. *Participants' task load* was manipulated by aircraft count in the right sector. In the 16 low task-load scenarios, there were four aircraft; in the 16 high task-load scenarios, there were ten aircraft. *Requestor's task load* was manipulated by aircraft count in the left sector. In the 16 low task-load scenarios, there were 2 aircraft; in the 16 high task-load scenarios, there were 8 aircraft. *Close landing demand* was manipulated by the call sign of the plane indicating their company. At the beginning of the experiment, participants were told that the drop-off gates of Air China (CA) and China Southwest Airlines (SZ) were nearer to the right runway under their control. The other two signs (MU and HU) were representing companies whose boarding gates were nearer to the left runway under the control of their colleagues. As a result, in the 16 no close-landing demand scenarios, the aircraft to be handover had the call-sign such as or; in the 16 close landing demand scenarios, the planes had the call sign CA or SZ.

Measuring Individual Difference Factors. *Work experience.* Work experience was measured by self-report. Participants indicated how long they have worked as a professional controller. *Job satisfaction.* Participants' job satisfaction was measured by the 5-item Job Satisfaction Scale [28]. Participant needed to rate to what extent they agreed with the statement such as "I find real enjoyment in my work" on a 5-point-Likert scale (1 for strongly disagree and 5 for strongly agree). The Cronbach alpha coefficient of the scale was .86.

2.3 Procedure

Upon arrival, participants first read and signed a written informed consent form. Then they completed a series of questionnaires about demographic information and personality (including Job Satisfaction). After that, they were asked to sit at a computer with a 22-inch-wide LED monitor and start to learn how to perform the task through 6 teaching scenarios. Once familiarized, they completed all 32 scenarios in a random manner.

3 Result

3.1 Initial Analysis

Table 1 presents the means, standard deviations, and correlations for all variables.

Table 1. Correlation matrix of all variables

| | Mean (SD) | 1 | 2 | 3 | 4 | 5 | 6 |
|---------------------------|-------------|-------|--------|-------|--------|-----|-----|
| 1. Close-landing | .50 (.50) | | | | | | |
| 2. Task-load | 7.00 (3.00) | – | | | | | |
| 3. Requestor's Task-load | 5.00 (3.00) | – | – | | | | |
| 4. Backing-up Willingness | 3.72 (2.75) | .05* | –.64** | .15** | | | |
| 5. Backing-up Decision | .54 (.49) | .08** | –.68** | .09** | –.72** | | |
| 6. Experience | 4.97 (3.93) | – | – | – | .02 | .06 | |
| 7. Job Satisfaction | 3.59 (.87) | – | – | – | .13 | .13 | .13 |

* $p < .05$, ** $p < .01$

3.2 HLM Analysis

To examine the effects of scenario parameters and individual difference simultaneously, avoiding problems caused by missing data, we adopted hierarchical linear modeling (HLM) with the HLM 6.02 program [29]. Using HLM, we tested our hypothesis in a nested approach. First, a null model with only the dependent variable was established. We found the intra-class correlation coefficient (ICC) was 12.7% for the willingness of backing-up, and 15.5% for the backing-up decision, respectively. According to [30], this indicates there was quite a large amount of between individual variance. Second, we tested the effects of task-level variables (level 1 predictors: Task-load, Requestor's Task-load, and Close Landing). Thirdly, we tested the main effects of the individual difference variables (level 2 predictors: work experience and Job Satisfaction) after controlling for the level predictors. Finally, we tested whether individual difference variables would interact with task-level variables.

HLM Results Predicting Backing-Up Willingness. In model 1 with task-level predictors, Task-load ($\beta = -.590$, $p < .01$), Requestor's Task-load ($\beta = .139$, $p < .01$), and Close Landing ($\beta = .315$, $p < .01$), turned out to be significant predictors of backing-up willingness. Controllers were more willing to accept the handover request when they were under low pressure, when their colleagues were at higher pressure and when the aircraft had a close-landing demand.

In model 2, the main effect of job satisfaction turned out to be a significant predictor ($\beta = .426, p < .05$), which means participants with higher job satisfaction were more willing to back up. Work experience, however, was not found to have a significant influence.

In model 3, it was found that Job Satisfaction had an interaction with Task-load ($\beta = -.078, p < .05$). We plotted such interaction in Fig. 2 to describe this moderating effect, based on the method recommended by [31]. It showed that Job satisfaction could improve controllers' backup intentions only when the task load is not high.

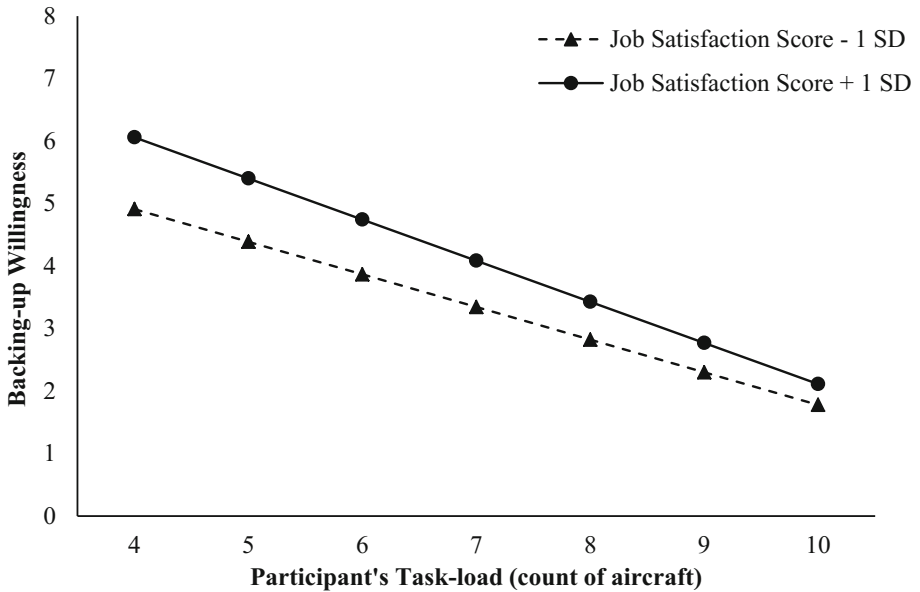


Fig. 2. The joint effect of Participant's Task-load and Job Satisfaction on Backing-up Willingness

HLM Results Predicting the Backing-Up Decision. In model 4 with task-level predictors, Task-load ($\beta = -.602, p < .01$), Requestor's Task-load ($\beta = .121, p < .01$), and Close Landing ($\beta = .624, p < .01$), turned out to be significant predictors of backing-up decisions. Controllers were more willing to accept the handover request when they were under low pressure, when their colleagues were at higher pressure and when the aircraft had a close-landing demand.

In model 5, the main effect of job satisfaction turned out to be a significant predictor ($\beta = .605, p < .01$), which means participants with higher job satisfaction were more willing to back up. Work experience, however, was not found to have a significant influence.

In model 6, it was found that Job Satisfaction had an interaction with Requestor's Task-load ($\beta = -.085, p < .01$). We plotted such interaction in Fig. 3 to describe this

moderating effect, based on the method recommended by [31]. It showed that Job satisfaction could improve the possibility of controllers' backup only when requestor's task load is not high (Fig. 3).

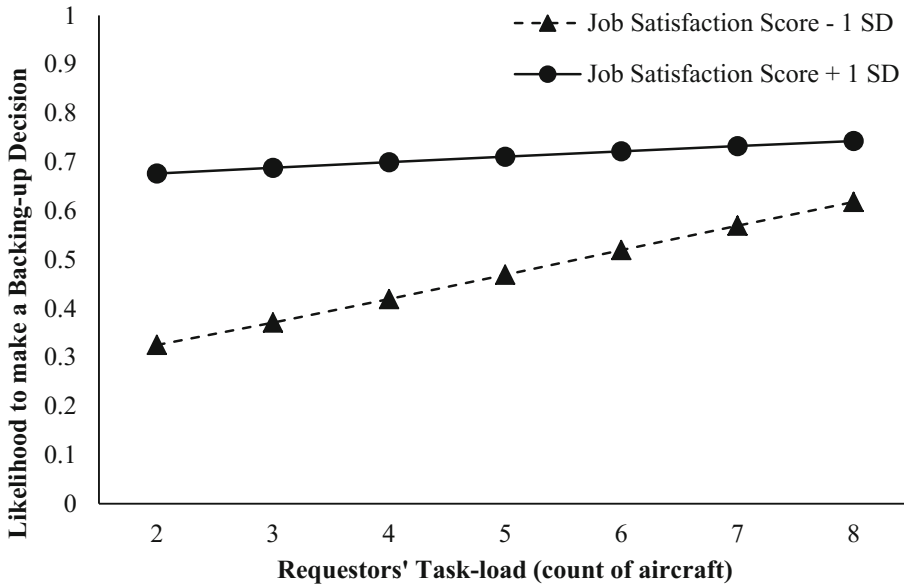


Fig. 3. The joint effect of Requestors' Task-load and Job Satisfaction on Backing-up Decision

Table 2. HLM results predicting Backing-up Willingness and Backing-up Decision

| Parameters | Backing-up Willingness | | | Backing-up Decision (1 for accept, 0 for reject) | | |
|--|------------------------|-------------------------|--------------------------|---|-------------------------|-------------------------|
| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
| Intercept | 3.716 (.16)** | 3.716 (.16)** | 3.716 (.15)** | .345 (.15)** | .366 (.19) ⁺ | .387 (.18) [*] |
| <i>Task level (N = 1280)</i> | | | | | | |
| Close Landing | .315 (.11)** | .315 (.10)** | .315 (.11)** | .624 (.16)** | .643 (.16)** | .652 (.16)** |
| Task-load | -.590 (.03)** | -.590 (.01)** | -.590 (.03)** | -.602 (.04)** | -.634 (.04)** | -.638 (.04)** |
| Requestor's Task-load | .139 (.02)** | .139 (.01)** | .139 (.02)** | .125 (.03)** | .129 (.03)** | .128 (.03)** |
| <i>Individual level (N = 40)</i> | | | | | | |
| Experience | | .004 (.04) | .004 (.03) | | .050 (.06) | .051 (.06) |
| Job Satisfaction | | .426 (.19) [*] | .426 (.16) [*] | | .605 (.19)** | .588 (.18)** |
| <i>Interaction</i> | | | | | | |
| Close Landing * Job Satisfaction | | | .195 (.10) | | | .057 (.16) |
| Task-load * Job Satisfaction | | | -.078 (.03) [*] | | | -.013 (.03) |
| Requestor's Task-load * Job Satisfaction | | | -.016 (.01) | | | -.085 (.02)** |

⁺ p < .10, ^{*} p < .05, ^{**} p < .01

4 Discussion

Focusing on the inter-sector cooperation between air traffic controllers, this study sought to explore how individual difference variables can influence controllers' backing-up behaviors. First, we confirmed previous findings that the three task-level variables, controllers' task-load, backup requestors' task-load, and the close landing demands, significantly predicted controllers' backing-up willingness and decision. Controllers were more willing to accept the handover request when they were under low pressure, when their colleagues were at higher pressure and when the aircraft had a close-landing demand. This can be explained in the framework that controllers evaluate their resources as well as the legitimacy of their team members' request [7–9].

Second, we found job satisfaction had a positive effect on both backup willingness and final acceptance decisions. Consistent with the literature of organizational citizenship behavior, controllers who had a higher level job satisfaction tended to provide more help to their colleagues beyond their task [21–24]. Moreover, we found that there was some interesting interaction between job satisfaction and specific task-level variables. When controllers' own task-load was high, all controllers were not willing to back up, regardless of their job satisfaction. However, when their task-load decreased, the help willingness of controllers with a higher level of job satisfaction increased more. This result suggests that people satisfied with their jobs were not acting like a “good old man” who will sacrifice their duties to help others. Also, when backup requestors' task-load was high, all controllers backed up a lot. And the backing-up behaviors of controllers with a higher level of job satisfaction decreased less when requestors' task-load decreased. While backup requestors' task-load was very high, it is such a legitimate request that almost no controllers refused to help. But when requestors' task-load was relatively low, requestor's situation was not so pressing, only controllers with higher job satisfaction still chose to back up, suggesting their behavior may not depend on overt cues of social legitimacy. Such effect is vital as whether a controller is in real need may not be easy to detect. A difficult air traffic situation may not always involve many aircraft, and few aircraft can also result in high demand. In this way, a timely backup that is not solely driven by the simple cues of workload (aircraft count) may be necessary for the controller as a whole.

Several limitations must be addressed before making any conclusion. First, although we designed a quite real environment and invited a quite large sum of professional controllers to participate our experiment, the study was still apparently a simulation, and the participants' behavior might be different from their behaviors in real work settings. Future studies may go further to collect real operational data which may depict the actual backup behaviors. Nevertheless, it is always a difficult decision to balance the need for ecological validity and adequate experimental control. Second, we only measured their job satisfaction in its natural form and the very nature of the design is cross-sectional. It is possible that a third unmeasured factor influences both job satisfaction and backup behaviors. Future studies may benefit from using intervention techniques such a job crafting or workspace design to improve controllers' positive feelings toward their job. These methods not only provide more robust evidence regarding a causal effect but also make actual improvement possible.

5 Conclusion

In this study, we found evidence suggesting that job satisfaction could be a significant predictor of controllers' backup behaviors. The effect of positive job attitude is new in the domain of ATC and has important implications. Since workload management is important for controllers, the traditional individual-focused approach might reach its limit. It is important to find new variables that may have a positive effect to promote workload redistribution. Future studies may find ways to cultivate a better attitude toward one's work.

Acknowledgements. This research was supported by National Key Research and Development Plan [grant number: 2016YFB1001203] and the Natural Scientific Foundation of China [31671148]. We thank the controllers and other working staffs that helped us to finish the experiment. E Xiaotian and Li Yazhe contributed equally to this paper.

References

1. ICAO: Global Air Transport Outlook to 2030 and trends to 2040 (No. Cir 333, AT, 190). ICAO, Montréal, Canada (2013)
2. Stankovic, S., Loft, S., Rantanen, E., Ponomarenko, N.: Individual differences in the effect of vertical separation on conflict detection in air traffic control. *Int. J. Aviat. Psychol.* **21**(4), 325–342 (2011)
3. Loft, S., Bolland, S., Humphreys, M.S., Neal, A.: A theory and model of conflict detection in air traffic control: incorporating environmental constraints. *J. Exp. Psychol.: Appl.* **15**(2), 106–124 (2009)
4. Neal, A., Kwantes, P.J.: An evidence accumulation model for conflict detection performance in a simulated air traffic control task. *Hum. Factors* **51**(2), 164–180 (2009)
5. 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)
6. Rantanen, E.M., Levinthal, B.R.: Time-based modeling of human performance. In: *Proceedings of the Human Factor and Ergonomics Society 49th Annual Meeting*, pp. 1200–1204. Sage Publications, Orlando (2005)
7. Loft, S.D., Humphreys, M.S., Neal, A.F.: Prospective memory in air traffic control. In: *Australian Aviation Psychology Symposium 2000*, vol. 1, pp. 287–294. Ashgate Publishing Company, Farnham (2003)
8. Porter, C.O.L.H., Hollenbeck, J.R., Ilgen, D.R., Ellis, A.P.J., West, B.J., Moon, H.: Backing up behaviors in teams: the role of personality and legitimacy of need. *J. Appl. Psychol.* **88**, 391–403 (2003)
9. Porter, C.O.L.H.: Goal orientation: effects on backing up behavior, performance, efficacy, and commitment in teams. *J. Appl. Psychol.* **90**, 811–818 (2005)
10. Barnes, C.M., Hollenbeck, J.R., Wagner, D.T., DeRue, D.S., Nahrgang, J.D., Schwind, K.M.: Harmful help: the costs of backing-up behavior in teams. *J. Appl. Psychol.* **93**(3), 529–539 (2008)
11. Norman, D.A., Bobrow, D.G.: On data-limited and resource-limited processes. *Cogn. Psychol.* **7**(1), 44–64 (1975)
12. Kahneman, D.: *Attention and Effort*. Prentice-Hall, Englewood Cliffs (1973)

13. Rouse, W.B., Edwards, S.L., Hammer, J.M.: Modeling the dynamics of mental workload and human performance in complex systems. *IEEE Trans. Syst. Man Cybern.* **23**(6), 1662–1671 (2002)
14. Bisseret, A.: Application of signal detection theory to decision making in supervisory control: the effect of the operator's experience. *Ergonomics* **24**(2), 81–94 (1981)
15. Sperandio, J.C.: Variation of operator's strategies and regulating effects on workload. *Ergonomics* **14**(5), 571–577 (1971)
16. Mogford, R.H., Guttman, J.A., Morrow, S.L., Kopardekar, P.: The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature. No. DOT/FAA/CT-TN95/22. Federal Aviation Administration, William Hughes Technical Center, Atlantic City, NJ (1995)
17. Laudeman, I.V., Shelden, S.G., Branstrom, R., Brasil, C.L.: Dynamic density: an air traffic management metric. No. NASA-TM-1988-11226. NASA Ames Research Center, Moffett Field, CA (1998)
18. Gianazza, D.: Forecasting workload and airspace configuration with neural networks and tree search methods. Elsevier Science Publishers Ltd. (2010)
19. E, X., Zhang, J.: Factors contributing to cross-sector back-up behaviors of approach controllers. Paper Presented at the First South Asia Region Conference of Psychology. Hanoi, Vietnam, November 2017
20. Ericsson, K.A., Charness, N., Feltovich, P.J., Hoffman, R.R.: *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, Cambridge (2006)
21. Bateman, T.S., Organ, D.W.: Job satisfaction and the good soldier: the relationship between affect and employee "citizenship". *Acad. Manag. J.* **26**(4), 587–595 (1983)
22. Williams, L.J., Anderson, S.E.: Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *J. Manag.* **17**(3), 601–617 (1991)
23. Rosenhan, D.L., Underwood, B., Moore, B.: Affect moderates self-gratification and altruism. *J. Pers. Soc. Psychol.* **30**(4), 546–552 (1974)
24. Clark, M.S., Isen, A.M.: Toward understanding the relationship between feeling states and social behavior. In: Hastorf, A.H., Isen, A.M. (eds.) *Cognitive Social Psychology*, pp. 71–108. Elsevier-North Holland, New York (1982)
25. E, X., Zhang, J.: Holistic thinking and air traffic controllers' decision making in conflict resolution. *Transp. Res. Part F: Traffic Psychol. Behav.* **45**, 110–121 (2017)
26. Zhang, J., Yang, J., Wu, C.: From trees to forest: relational complexity network and workload of air traffic controllers. *Ergonomics* **58**(8), 1320–1336 (2015)
27. Zhang, J., Du, F.: Relational complexity network and air traffic controllers' workload and performance. In: Harris, D. (ed.) *EPCE 2015. LNCS (LNAI)*, vol. 9174, pp. 513–522. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20373-7_49
28. Judge, T., Lock, E., Durham, C., Kluger, A.: Dispositional effects on job and life satisfaction: the role of core evaluations. *J. Appl. Psychol.* **83**(1), 17–34 (1998)
29. Raudenbush, S.W., Bryk, A.S.: *Hierarchical Linear Models: Applications and Data Analysis Methods*, vol. 1. Sage Publications, Orlando (2002)
30. Peugh, J.L.: A practical guide to multilevel modeling. *J. Sch. Psychol.* **48**(1), 85–112 (2010)
31. Frazier, P.A., Tix, A.P., Barron, K.E.: Testing moderator and mediator effects in counseling psychology research. *J. Couns. Psychol.* **51**(1), 115–134 (2004)



Quantitative Study of Alertness During Continuous Wakefulness Under the Effect of Nervous Activity

Kang Li, Ruishan Sun^(✉), Jingqiang Li, and Yu-Ting Zhang

Research Institute of Civil Aviation Safety, Civil Aviation University of China,
Tianjin, China

kangkanggo@outlook.com, sunrsh@hotmail.com,
ljqtianjin@126.com, zhangyt0715@outlook.com

Abstract. To examine changes of pilot's alertness during lengthy periods of continuous wakefulness, we used a Nervous System Assessment Tool to assess fluctuations in subjects' reaction times under the effect of different nervous states: nervous excitation, nervous inhibition, and nervous stability. And we followed used Polynomial Fit to quantitatively assess the reaction time. Results showed that the responses of the three neurological states are significantly affected by circadian and homeostatic drives. Under the effect of nervous inhibition, reaction time enters the state of fatigue earlier and shows a longer period of decline; under the effect of nervous stability, reaction time enters the state of fatigue later, and the period of decline in alertness is shorter. The experimental results are meaningful for the rostering of pilots.

Keywords: Alertness · Reaction time · Nervous activity · Sleep deprivation
Pilots

1 Introduction

Alertness refers to the ability of an organism to maintain attention and remain vigilant during long periods of time [23]. Maintaining alertness in personnel is a vital part of ensuring safety in production, especially in the field of aviation, in which flight safety must be guaranteed at all times. The importance of alertness in aviation has been demonstrated by numerous flight accidents, such as the Guantánamo air disaster (1993), the Korean Air Flight 801 disaster (1997), and the Air Berlin pan pan distress call made because of pilot fatigue while approaching Munich, Germany (2012). Caldwell et al. considered that the short sleep time, the early awakening and jet lag could reduce the alertness, flight ability and lead to fatigue of pilots [8]; Blakey also pointed out that sleep time, circadian rhythm, and sleep quality affect the alertness of pilots [17]. Sleep time will be taken up by flight duration and caused fatigue of pilots, especially the long-range pilots: long-range flight have longer flight duration, combined with the effect of jet lag and cabin conditions (such as: light, temperature, noise, turbulence and so on), make pilots to stay awake for a long time. For example, 25 of the 392 long-range flights from Singapore to New York reported no sleep record of pilots

in 2005 [1]. It is obvious that quantizing the alertness of pilots, studying its change rule during continuous wakefulness, and undertaking necessary measures to deal with decreased alertness are very important in increasing aviation safety and decreasing accident rates.

Borbe'ly made early studies of the level of alertness of human body over the course of a day, and he thought that alertness depends on homeostatic and circadian drives of the body [18]. Monk et al. discussed the factors influencing human performance by measuring the secretion of human hormones, arguing that circadian rhythms have an impact on job performance [19]. Akerstedt and Folkard proposed the Three-Process Model of Alertness, which elaborates further on the role of circadian and homeostatic mechanisms in alertness under conditions of continuous conscious wakefulness [3]. Circadian rhythms refers to the rhythms of human physiological activities, and they are an evolutionary mechanism that enable the human body to adapt to long-term choices of environment [2]. They are usually expressed as a 24-h sinusoidal curve; however, because of individual differences, there is no exact representation. They generally show a rise from 6:00 a.m. to 18:00 p.m., peaking at 18:00, and declining from 18:00 to 6:00 [12]. The homeostatic mechanism is closely related to the amount of sleep, and it maintains alertness. Its value is lowest during the wake-up phase, and it gradually increases as time passes after waking up [3].

In the early days of alertness assessment, physiological signals, including blink frequency, skin resistance, body temperature, and blood pressure were used to evaluate alertness [7]. As research methods advanced, alertness assessment focused mainly on the degree of fatigue and subjective consciousness, reaction time, EEG activity, and hormone secretion [9]; e.g., Jung et al. assessed the state of alertness of operators through EEG measurements [13]; Nicholson et al. used EEG measurements to study the night sleep and alertness of a flight crew flying from London to San Francisco [20]; Badia et al. confirmed that alertness is regulated by the circadian rhythm system according to the beta activity of the brain and the body temperature index [4]. Biochemical and physiological signal detection methods have clear indicators and accurate test results, but rigorous testing conditions, complex testing process, and research results are not easy to generalize and apply [10]. Compared to other testing methods, subjective evaluation and bioreaction testing are much more convenient, rapid, and effective approaches for assessing the alertness of personnel. Sallinen et al. applied the Karolinska sleepiness scale (KSS) to evaluate both sleep and duty alertness in long-haul airline transport pilots [22]. The psychomotor vigilance test (PVT), which is an assessment tool based on reaction time, is also used widely by researchers [5, 16, 24]. The above researches mainly focus on the measurement of instantaneous alertness and fail to carry out the alertness analysis based on the information processing ability of the human brain. During the working process of first-line staff such as pilots, the brain needs certain information processing to maintain the work performance, making the above studies have some limitations in the practical application process.

In addition, with the process of brain information processing, nervous activity characteristics will produce a corresponding change [6], and according to the theory of neural activity, neuronal cells in the cerebral cortex have the characteristics of excitation and inhibition, and a combination of changes in these characteristics affects alertness [21]. Different people have different nervous activity and enter the state of

fatigue differently [11, 25]. Early foreign scholars used the “Uchida–Klinebrene measurement method” and “Amphim scales” and other methods to measure changes in nerve type, and domestic scholars have also designed the 80.8 Nervous System type measurement to assess the types of nerves in the process of brain information processing [26, 27]. However, such studies mainly focus on the measurement of nervous activity characteristics, failed to be linked with the alertness in the production process, and the maintenance of alertness is of great significance to ensure safety in aviation.

Our study uses the nervous system assessment method to study changes in the alertness of subjects with the process of brain information processing in a 36 h state of wakefulness, more deeply from the perspective of neural function and hope to provide new ideas for the fatigue management of pilots.


2 Materials and Methods

2.1 Subjects

The subjects in this study were 6 young students (22–25 years of age), with an average age of 23.6 years. They were right-handed, and in good health, with healthy lifestyles; e.g., they did not smoke or drink alcohol, did not drink coffee, did not take drugs, and maintained a regular routine. All subjects gave their written informed consent for the study, and the experiment was approved by the Ethics Committee of Civil Aviation University of China.

2.2 Equipment

Nervous System Assessment of the aviation personnel safety risk assessment system involves a fatigue-inducing task and the use of recording equipment to measure reaction time. The Nervous System Assessment Tool is a crossing-out experiment that references the Uchida–Kraepelin test for psychological stress [26], 80.8 Nervous System type measurement [27], and the BTL-QZ Test [15] and is based on the visual–action conditional reflex and performance testing principles. In the course of experimental operation, the subjects’ nervous system activities are continuously in a state of conversion of synaptic excitation to inhibition. According to Pavlov’s theory, the operation results are divided into three dimensions: excitation, inhibition, and stability (excitement–inhibition) of the nervous system, which can be realized through a crossing-out experimental design.

Figure 1 shows the visual test interface of software programming; the interface is divided into three areas: (1) The Reading Area consists of 10 symbols and is divided into 6 rows, each of which consists of 20 symbols freely combined, making a total of 120 symbols {  }; (2) The Marking











Area consists of 2 target symbols (these appear at random and are represented by the following symbols: {        }) and a Judgment symbol { }; (3) The Operating Area includes three basic operations: Circle, Stroke, and Ignore [14].



Fig. 1. Test interface

The operating rules are shown in Table 1.

Table 1. Operating rules

| Previous symbol | Current symbol | |
|---|------------------------|--|
| | Target symbol α | Non-target symbol $\underline{\alpha}$ |
| Judgment symbol β | Circle C | Ignore I |
| Non-judgment symbol $\underline{\beta}$ | Stroke D | Ignore I |

The process of the operation rules designed by the tester is shown in Fig. 2. According to the distribution and meaning of theoretical and practical results,

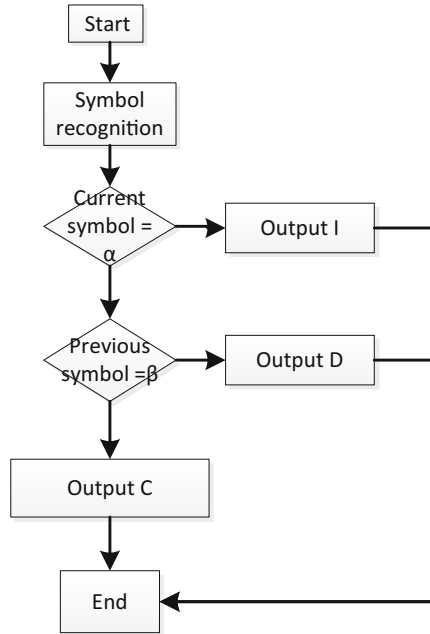


Fig. 2. Operating rules' flow chart

measurement results is defined as follows: D/D-R is the reaction time under nervous excitation; C/C-R is the reaction time under nervous stability; and I/I-R is the reaction time under nervous inhibition. This tool has been demonstrated in the correlation between human error and the characteristics of nervous system activity [14].

2.3 Procedures

Six subjects were asked to ensure that they had normal sleep the day before the experiment and to enter the laboratory at 8 am on the first day for continuous sleep deprivation (SD) of 36 h. The reaction time data collected during the 36 h were used to evaluate the status of alertness of the participants during continuous wakefulness. During the 36 h, the participants took a unit of 1 h as the assessment cycle. They simulated the control task for 40 min, and the experimental data were collected for 6 min, and the remaining 14 min were used to adjust the distribution. During the experiment, there was constant supervision by the managers to prevent participants either from touching any stimuli or falling asleep.

2.4 Analytical Approach

The database was established by using spss22.0, and descriptive statistics and ANOVA were also carried out to analyze the characteristics of and differences between the three

types of reaction time. We used OriginPro 2017 to make drawings and conducted Polynomial Fit between sleep deprivation time and reaction time.

Alertness analysis was based on the reaction time in 36 h sleep deprivation time and was assessed using three different ways: Fluctuation trend in 36 h of 6 subjects' reaction time by the way of drawing Grouped Box Charts; The difference of reaction time in three sleep deprivation time period (1–12 h, 12–24 h, 24–36 h) by the way of analyzing mean, standard deviation and ANOVA, and the result was shown by mean \pm standard deviation in table; Mathematical relations between reaction time and sleep deprivation time by Polynomial Fitting.

3 Results

As shown in Figs. 3, 4 and 5, the Grouped Box Charts were made using 3 h time-periods to show the reaction time of the six participants under the three types of nervous effect over 36 h. Ignoring the learning effect in the first three hours (the first time period), it can be seen from the figure that the reaction time fluctuates with time and that there are abnormal values with a large degree of deviation.

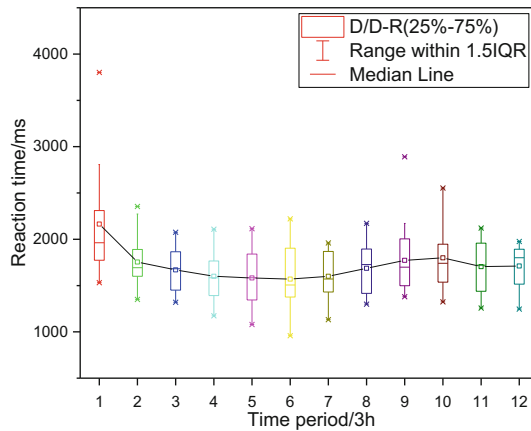


Fig. 3. D/D-R of 6 subjects during 12 time periods

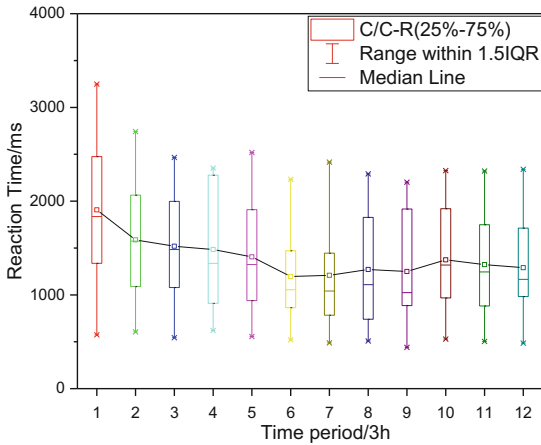


Fig. 4. C/C-R of 6 subjects during 12 time periods

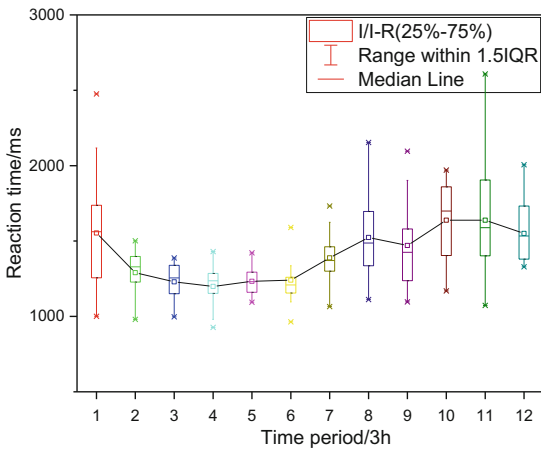


Fig. 5. I/I-R of 6 subjects during 12 time periods

From the result of mean (as shown in Table 2), three kinds of reaction time have the lowest reaction in 12–24 h; compared with the two time period between 1–12 h and 24–36 h, the D/D-R and C/C-R decreased, while I/I-R increased after 24 h sleep deprivation. According to the results of standard deviation, the discrete degree of three kind reaction time were decrease by the time. Through the descriptive statistics and ANOVA of the mean of D/D-R, C/C-R and I/I-R of the six subjects (as shown in Table 3), we found that D/D-N, C/C-N, I/I-N, C/C-R and I/I-R were significantly different in three time periods at the confidence level of 0.01; D/D-R was significantly different in three time periods at the confidence level of 0.05.

Table 2. Results of descriptive statistics and ANOVA in three sleep-deprivation time-periods

| SD/h | 1–12 | 12–24 | 24–36 | F | Sig |
|-------|------------------|------------------|------------------|-------|-------|
| D/D-R | 1796.75 ± 276.43 | 1609.33 ± 66.67 | 1746.75 ± 78.27 | 3.89 | 0.030 |
| C/C-R | 1624.25 ± 221.82 | 1270.75 ± 115.42 | 1309.33 ± 80.05 | 19.64 | 0.000 |
| I/I-R | 1318.08 ± 183.48 | 1346.42 ± 140.21 | 1574.42 ± 103.79 | 11.09 | 0.000 |

According to the Box Chart distribution characteristics of each indicator, the greater dispersion degree data (outlier) of the six subjects are eliminated. The Polynomial Fit was made by using the mean of D/D-R, C/C-R and I/I-R of the six subjects as the dependent variable and the SD time as an independent variable. The Parameters, Statistics and ANOVA of the Polynomial Fit are shown in Tables 3 and 4, and the Fitted Curves Plot is shown in Fig. 6.

Table 3. Parameters of Polynomial Fit

| Y | | Value | t-value | P |
|-------|-----------|----------|---------|-------|
| D/D-R | B1 | -103.032 | -10.534 | 0.000 |
| | B2 | 5.636 | 10.007 | 0.000 |
| | B3 | -0.087 | -9.181 | 0.000 |
| C/C-R | Intercept | 2149.563 | 45.159 | 0.000 |
| | B1 | 65.253 | 1.759 | 0.046 |
| | B2 | -10.941 | -3.075 | 0.003 |
| | B3 | 0.480 | 3.606 | 0.000 |
| | B4 | -0.006 | -3.808 | 0.000 |
| I/I-R | Intercept | 1498.667 | 12.298 | 0.000 |
| | B1 | -89.719 | -5.438 | 0.000 |
| | B2 | 5.814 | 6.119 | 0.000 |
| | B3 | -0.093 | -5.818 | 0.000 |
| | Intercept | 1596.763 | 19.886 | 0.000 |

Table 4. Statistics and ANOVA of Polynomial Fit

| | Polynomial order | F-value | P | R ² |
|-------|------------------|---------|-------|----------------|
| D/D-R | 3 | 43.831 | 0.000 | 0.814 |
| C/C-R | 4 | 33.548 | 0.000 | 0.822 |
| I/I-R | 3 | 59.221 | 0.000 | 0.856 |

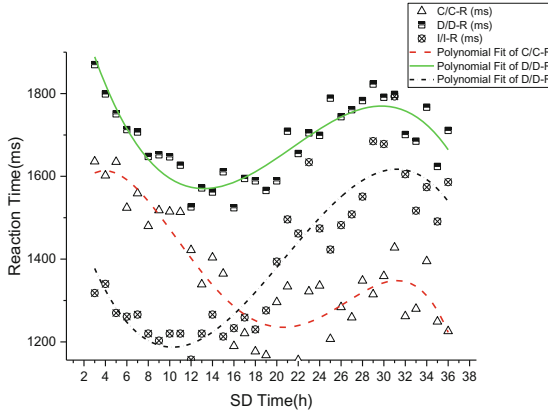


Fig. 6. Fitted curves plot of reaction time

The fluctuation trend of each indicator in 36 h is shown in the Fig. 6.

The Fitting result between C/C-R and SD can explain 82.2% of the variation rate, which was significant according to ANOVA ($F = 33.548, P < 0.01$). The fitting model is:

$$Y = 1498.667 + 65.253x - 10.941x^2 + 0.480x^3 - 0.006x^4;$$

The Fitting result between D/D-R and SD can explain 81.4% of the variation rate, which was significant according to ANOVA ($F = 42.831, P < 0.01$). The fitting model is:

$$Y = 2149.563 - 103.032x + 5.636x^2 - 0.087x^3;$$

The Fitting result between I/I-R and SD can explain 85.6% of the variation rate, which was significant according to ANOVA ($F = 59.221, P < 0.01$). The fitting model is:

$$Y = 1596.763 - 89.719x + 5.814x^2 - 0.093x^3;$$

Excluding the learning effect on the reaction time of first two groups, there were differences among the responses of the three neurological fluctuations during the 36 h. D/D-R showed the highest reaction time, while the difference in reaction times between I/I-R and C/C-R was small. D/D-R increased from 8:00 to 19:00 under the influence of circadian rhythms and showed the slowest reaction time around 19:00. Under the effect of homeostatic driving and circadian rhythms, D/D-R showed the fastest rate of increase from 12:00 to 6:00, with the upward trend disappearing around 15:00 the next day. With the increase of circadian rhythms after 15:00, D/D-R decreased to a certain degree. I/I-R will be earlier into the state of fatigue: from 8:00 to 17:00, I/I-R decreased. Compared to D/D-R, which showed its slowest reaction time at about 19:00, I/I-R

reached its minimum reaction time at 17:00; after 17:00 the I/I-R increased and the upward trend disappeared around 15:00 the next day. Thus, the period of increase of I/I-R is longer than that of D/D-R. C/C-R will be later into the state of fatigue: C/C-R decreased from 10:00 to 3:00. Compared to D/D-R, which reached its slowest reaction time at about 19:00, I/I-R reached its slowest reaction time at about 3:00 the next day. Thus, the period of increase is shorter for I/I-R.

4 Discussion

Previous researches on the alertness of pilots were mostly based on treatise or interviewing pilots [8, 17], even though some researches use the measure of experiments, the data of which is just about pre-flight and post-flight [5, 20, 22], that will miss the alertness of pilots during the specific implementation of the flight, and the corresponding fatigue management measures lack persuasion, can not be specific and detailed guidance flight mission. Our research study the alertness of subjects during continuous 36 h, can effectively make up the lack of previous researches, and for the first time from the perspective of nervous activity to study alertness of the pilots, can propose more innovative and effective pilot management measures.

The experimental results confirm that circadian and homeostatic driving have an influence on the alertness of personnel [3, 18, 19]. Especially in the late night from 3:00–6:00, when the circadian and homeostatic drive have negative impact on the alertness at the same time: the reaction time of three nervous activity all have the fastest rate of rise, and subjects are difficult to focus and maintain the work performance during this period in practical work. What's more, the time under different nervous activity enter the state of fatigue are different. According to the results of the experiment, alertness under the effect of nervous inhibition maintain the state of conflict (which can be confirmed by the larger slope of the curve and the large difference between the front and back time periods in Fig. 6), with a earlier decline, and shorter time into the state of fatigue. Under the effect of nervous inhibition, subjects unable to maintain work performance for a long time. Nervous stability can reflect the stability of the individual [15], and according to the fitted curve results (Fig. 6), there is a steady trend and lower differences in the later time period, indicating that nervous stability have stronger toughness compare to other types of nerves and a longer time to maintain alertness. Under the effect of nervous stability, subjects able to maintain a longer working ability.

The application of current research findings might be able to integrate to airlines safety management systems for pilots' rostering. For example: the pilots have a stronger nervous stability can have a stronger ability to continue working in a long time who can be selected for long-range flight and night flight, while the pilots have a stronger nervous inhibition who will be earlier into the state of fatigue and have a suddenly alert dropped in night, are not suitable for long hours of continuous flight, should be avoided carrying out long-range flight and night flight. In general, pilots should be arranged night shift as few as possible, they are difficult to focus and maintain performance in night time; If night work can not be avoided, pilots have to ensure adequate sleep before performing night shifts, which can greatly help to

improve alertness [8]; And multiple pilots should be set to strengthen the work rotation in long-range flight, to avoid a sharp decline of alertness on the ability to fly in long awakening time.

This study focuses on the change of the alertness of the general public during continuous wakefulness and does not divide according to the types of circadian rhythms. However, the alert performance of pilots with different rhythmic types will be different [2]. Subsequent studies should refine the impact of rhythmic patterns on alertness and provide more detailed advice on flight fatigue management.

5 Conclusion

Based on a 36 h sleep-deprivation experiment, reaction time under the influence of nervous excitation, nervous inhibition, and nervous stability were quantitatively analyzed by Polynomial Fit. We drew the following conclusions: (1) Alertness under different nervous activities is significantly affected by the circadian and homeostatic drives. In general, alertness increases from the point of morning and reaches its highest point around nightfall. It declines most rapidly at midnight and minimum alertness is reached around 15:00 the next day. (2) Alertness varies under different states of nervous activity: under nervous inhibition, reaction time enters the fatigue state earlier and shows a longer period of decline, and with a weaker ability to maintain work performance in a long time; under nervous stability, reaction time enters the state of fatigue later, and the period of decline in alertness is shorter, and it with a stronger ability to maintain work performance for a long time. This study provides a new perspective of measuring alertness and of managing and controlling pilots fatigue to some extent, and the study may help the pilots' rostering according to the difference of nervous activity which is meaningful for the safety of aviation.

Acknowledgements. The financial support of the Humanity and Social Science Youth Foundation of Ministry of Education of China under grant number 15YJC190008. This work also supported in part by the Fundamental Research Funds for the Central Universities of China under grant numbers 3122017014.

References

1. Flight Safety Foundation, Incorporated: Lessons from the dawn of ultra-long-range flight. *Flight Safety Digest*, vol. 24 (2005)
2. Abbott, S.M., Reid, K.J., Zee, P.C.: Circadian rhythm sleep-wake disorders. *Psychiatr. Clin. N. Am.* **38**(4), 805 (2015)
3. Åkerstedt, T., Folkard, S.: Predicting sleep latency from the three-process model of alertness regulation. *Psychophysiology* **33**(4), 385–389 (1996)
4. Badia, P., Myers, B., Boecker, M., Culpepper, J., Harsh, J.R.: Bright light effects on body temperature, alertness, EEG and behavior. *Physiol. Behav.* **50**(3), 583 (1991)
5. Basner, M., Mollicone, D., Dinges, D.F.: Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta Astronaut.* **69**(11–12), 949–959 (2011)

6. Borisyyuk, R., Chik, D., Kazanovich, Y.: Partial synchronization of neural activity and information processing. In: International Joint Conference on Neural Networks, vol. 21, pp. 3399–3406 (2009)
7. Cajochen, C., Khalsa, S.B., Wyatt, J.K., Czeisler, C.A., Dijk, D.J.: EEG and ocular correlates of circadian melatonin phase and human performance decrements during sleep loss. *Am. J. Physiol.* **277**((3 Pt 2)), R640 (1999)
8. Caldwell, J.A., Mallis, M.M., Caldwell, J.L., Paul, M.A., Miller, J.C., Neri, D.F.: Fatigue countermeasures in aviation. *Aviat. Space Environ. Med.* **80**(1) (2009)
9. Chang, A.M., Scheer, F.A., Czeisler, C.A., Aeschbach, D.: Direct effects of light on alertness, vigilance, and the waking electroencephalogram in humans depend on prior light history. *Sleep* **36**(8), 1239 (2013)
10. Gao, Y., Li, H., Li, J.Q.: The research of the influence of awakening time and sleep length to the vigilance of control post. *Sci. Technol. Eng. (Chin.)* **16**(36), 147–151 (2016)
11. Guo, Z.L., Meng, P.X., Zheng, R.C.: Development of the Pavlovian temperament survey Chinese edition (PTS-C). *Chin. Mental Health J. (Chin.)* **18**(2), 91–93 (2004)
12. Hursh, S.R.: System and method for evaluating task effectiveness based on sleep pattern. US6579233 (2003)
13. Jung, T.P., Makeig, S., Stensmo, M., Sejnowski, T.J.: Estimating alertness from the EEG power spectrum. *IEEE Trans. Biomed. Eng.* **44**(1), 60 (1997)
14. Li, J.Q., Li, H., Wang, Y., Zhao, N.: Research on correlations between human errors of raw controllers and characteristics of their nervous system. *China Saf. Sci. J. (Chin.)* **27**(3), 13–18 (2017)
15. Li, L.: Revision on temperament measurement system of BTL-QZ-V 1. *Psychol. Explor. (Chin.)* **28**(2), 85–90 (2008). 1 edition
16. Lim, J., Ebstein, R., Tse, C.Y., Monakhov, M., Lai, P.S., Dinges, D.F., et al.: Dopaminergic polymorphisms associated with time-on-task declines and fatigue in the psychomotor vigilance test. *Plos One* **7**(3), e33767 (2012)
17. Loewenthal, K.M., Eysenck, M., Harris, D., Lubitsh, G., Gorton, T., Bicknell, H.: Stress, distress and air traffic incidents: job dysfunction and distress in airline pilots in relation to contextually-assessed stress. *Stress Health* **16**(3), 179–183 (2015)
18. Marzano, C., Ferrara, M., Curcio, G., Gennaro, L.D.: The effects of sleep deprivation in humans: topographical electroencephalogram changes in non-rapid eye movement (NREM) sleep versus REM sleep. *J. Sleep Res.* **19**(2), 260–268 (2010)
19. Monk, T.H., Buysse, D.J., Rd, R.C., Berga, S.L., Jarrett, D.B., Begley, A.E., et al.: Circadian rhythms in human performance and mood under constant conditions. *J. Sleep Res.* **6**(1), 9–18 (1997)
20. Nicholson, A.N., Pascoe, P.A., Spencer, M.B., Stone, B.M., Green, R.L.: Nocturnal sleep and daytime alertness of aircrew after transmeridian flights. *Aviat. Space Environ. Med.* **57** (2), 43–52 (1986)
21. Regan, D.: Nervous activity. *Nature* **324**(6095), 310–311 (1986)
22. Sallinen, M., Sihvola, M., Puttonen, S., Ketola, K., Tuori, A., Härmä, M., et al.: Sleep, alertness and alertness management among commercial airline pilots on short-haul and long-haul flights. *Accid. Anal. Prev.* **98**, 320–329 (2017)
23. Shaw, T.H., Matthews, G., Warm, J.S., Finomore, V.S., Silverman, L., Costa Jr., P.T.: Individual differences in vigilance: personality, ability and states of stress. *J. Res. Pers.* **44** (3), 297–308 (2010)
24. Sheldon, S.H., Ferber, R., Kryger, M.H.: Principles and practice of pediatric sleep medicine. *Archiv. Dis. Child.* **91**(6), 546 (2006)

25. Smith, O.R., Pedersen, S.S., Van Domburg, R.T., Denollet, J.: Symptoms of fatigue and depression in ischemic heart disease are driven by personality characteristics rather than disease stage: a comparison of CAD and CHF patients. *Eur. J. Cardiovasc. Prev. Rehabil.: Official J. Eur. Soc. Cardiol. Working Groups Epidemiol. Prev. Cardiac Rehabil. Exerc. Physiol.* **15**(5), 583–588 (2008)
26. Sugimoto, K., Kanai, A., Shoji, N.: The effectiveness of the Uchida-Kraepelin test for psychological stress: an analysis of plasma and salivary stress substances. *Biopsychosocial Med.* **3**(1), 5 (2009)
27. Zhang, Q.H.: Study on the measurement of human nerve type. High Education Press (Chinese) (1993)



Tracking Provenance in Decision Making Between the Human and Autonomy

Crisrael Lucero^(✉), Braulio Coronado^(✉), Eric Gustafson^(✉),
and Douglas S. Lange^(✉)

Space and Naval Warfare Systems Center Pacific, San Diego, USA
{crisrael.lucero,braulio.coronado,eric.a.gustafson1,doug.lange}@navy.mil

Abstract. Provenance has been used as a measure of accountability, trust, and validity. Within the context of Command and Control (C2), provenance can be utilized to track the decision making processes that change data and dependencies. C2 of several unmanned autonomous vehicles provide a complex and ever-changing battlespace, with many actors and decision makers. The goal of this paper is to track and explain the autonomous decisions made by an intelligent C2 station based on its interactions with various systems using provenance. With provenance providing explanations and reasoning behind the actions of autonomy, a form of system accountability and transparency is achieved between human and machine.

Keywords: Provenance · Autonomous systems · Decision making
Situation awareness

1 Introduction

Provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or thing [1]. Provenance explains more than what happened, but it also answers how data was manipulated, why, and who was involved in the process. Data Provenance explains the interactions of these concepts within a system and has the potential to aid the decision making process in a complex battlespace with many autonomous actors. The PROV Data Model, PROV-DM, is a generic data model for the standardized open provenance model described by W3C [2]. Fully recorded provenance will illuminate dependencies, responsibility flow, and explain why certain actions were made. Dependency and decision tracking is a critical piece of C2, especially within a battlespace of several autonomous actors, whether human or machine.

We investigate tracking decisions made and actions taken between the human and autonomy within a prototype centralized C2 system, the Intelligent Multi-UxV Planner with Adaptive Collaborative Control Technologies (IMPACT),

The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.

using provenance. In this study, we seek to track provenance of decision makers, whether software agents or human operators, in order to derive responsibility flow within IMPACT. Afterwards, we describe ideas for future work and avenues to take with the information learned (Fig. 1).

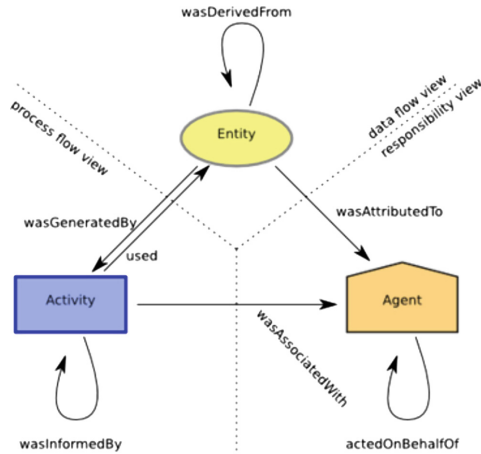


Fig. 1. Visual representation of how data provenance is modeled. This figure showcases the three different views alongside the traditional PROV layout convention. Entities are represented as yellow ellipses, activities are blue rectangles, and agents are orange pentagon-houses. Used by permission of the author [1] (Color figure online)

2 Provenance and Decision Making

Studies have mostly done assessments on provenance to determine quality, reliability, or trustworthiness of data using offline provenance, or provenance that is generated after applications have finished running. We are taking an approach to dynamically record and generate provenance during system runtime, or online provenance as defined by Sezavar Keshavarz et al. [3]. Further research with online provenance has shown that online provenance have aided online decision making and behavior recognition of crowdsourced activities were derivable [4]. Accountability is a critical piece of autonomy within C2; provenance has the ability to trace decisions that change information made by autonomous systems back to operators and vice versa.

Although Keshavarz's study showcases online provenance for decision making for computational systems, the information derived from the methodology can be translated and passed on to human decision makers. Conversely, when a decision can be considered as the provenance of the selected course of action, it allows human decision management to be translated into a machine-understandable representation of decisions using provenance [5]. Provenance can provide both

transparency of a system to a human while also allowing a machine to perform analytics based on the history of actions by humans.

The PROV Data Model allows for domain and application specific representations of provenance that was designed to be translated and interchanged between systems. Different heterogeneous systems can export their native provenance into the PROV-DM and the model allows for applications to make sense of the provenance and reason over it [2]. This is one of the strengths provided by PROV-DM that opens the doors to human-computer-collaboration.

3 Architecture

3.1 IMPACT

IMPACT is a C2 prototype platform for centralized supervised control of simulated autonomous unmanned vehicles [6]. The underlying goal of the IMPACT system is to allow an inversion of the unmanned vehicle staffing ratio. Currently, many operators are required to control a single unmanned vehicle [7]. IMPACT explores the use of higher platform and application level autonomy to grant a single operator control of teams of disparate unmanned vehicles. Through a “playbook” approach, an operator calls a “play” which tasks a vehicle or team of vehicles with mission objectives [8]. A play can involve multiple vehicles such as an overwatch play, where the vehicle being watched and the vehicle performing the watching can be of completely different types (e.g. an air vehicle watching a surface vehicle which is itself performing a point inspect play). Furthermore, operators can configure plays with constraints requiring certain payloads to accomplish the task and can be explicit or implicit (e.g. require a certain payload, or set environment constraints that implicitly require a certain payload).

Data driving this prototype flows through a centralized hub, where modules specializing in a certain function subscribe to messages required for their operation. An autonomous agent subscribes to and processes play requests to allocate resources and recommend assets based on the current state of all available assets combined with the constraints specified by the operator. An automated route planner ingests play requests and generates routes based on current vehicle, environment and mission requirements. There are three modules of interest residing within IMPACT that are suitable for Provenance instrumentation:

Fusion. The Fusion agent realizes visualization of the combined IMPACT modules. This Human Computer Interface (HCI) acts as the primary means of interaction between the human operator and simulated vehicles. Fusion presents a high-level C2 perspective with tools enabling play-calling, environment editing and interfaces for the various modules composing IMPACT. The “playbook” visually represents plays available for the operator and provides an interface to specify play constraints and asset allocation preferences. Upon calling a play, Fusion will display routes and task destinations for vehicles performing the play.

Task Manager. The Task Manager (TM) aids the operator by helping to manage workload. Previous work has investigated the use of autonomous task management in order to balance operator attention with situational awareness in the unmanned vehicle domain [8]. Certain events in IMPACT trigger the generation of tasks within the TM for the operator to perform. In addition to events, chat conversations are parsed for relevant content that may lead to the generation of new tasks. As tasks are generated in TM, operator-configured working agreements allocate tasks based on what the operator has decided the autonomy is allowed to do. This helps the operator achieve more control and builds trust in the autonomy as it guides the behavior of the autonomy towards operator expectations. Additionally, some generated tasks are pre-configured to map directly to certain plays, with the parameters necessary to accomplish the play pre-filled by the TM with the goal of increasing the speed of the play-calling process (Fig. 2).



Fig. 2. The Task Manager is populated with predicted tasks based on chat messages.

Plan Monitor. The Plan Monitor (PM) observes ongoing missions and relays plan health evaluation to the operator. The PM generates a formal network model of entities and components in IMPACT within the Rainbow Autonomics Framework [6]. This model is primed with thresholds and constraints ensuring mission plan requirements and vehicle status are within specifications. In addition to mission plan health, constraints on vehicle properties ensure vehicle status is monitored. Evaluation of the model happens near real time and adaptation strategies are triggered upon model constraint violations. Plan health is represented in the PM tile within the Fusion interface and is color coded to visually depict mission plan quality. Vehicle status is communicated through alert messages presented as a text overlay. Finally, PM interacts with a policy

checking module, described below, that allows it to directly and autonomously modify ongoing plans under specific conditions.



Fig. 3. The Plan Monitor displays plan health for active plays based on several criteria such as vehicle energy and time to destination.

3.2 COMPACT

The Configurable Operating Model Policy Automation for Control of Tasks (COMPACT) system is a C2 policy compliance checking system for unmanned air vehicles (Fig. 3). COMPACT performs policy compliance information checks and sends the information to IMPACT in order for the application level autonomy to reroute ongoing missions, update new restricted or keep-in zones, and call plays to maintain policy compliance. This autonomous reroute play calling is achieved under specific conditions and through communication with a trio of modules: COMPACT, PM and TM. Upon detection of an air vehicle projected to enter a restricted zone, COMPACT publishes a predicted air space violation alert which is ingested by PM. PM updates its model and triggers an adaptation strategy to publish a task generation message which is received by TM. Finally, TM generates a reroute mission task notifying the operator that a vehicle is predicted to enter a restricted area. If a TM working agreement is configured such that the autonomy is allowed to reroute plans, the task is executed immediately.

3.3 MAPLE

The Maritime Autonomous Platform Exploitation (MAPLE) framework is a Ground Control Station (GCS) with several modules used for remote control of unmanned vehicles. Systems within the MAPLE framework allow operators and

pilots to fly actual vehicle assets, which allows IMPACT to cross the boundary between simulated and live unmanned vehicles. Human approval on the MAPLE side is required for all missions planned by IMPACT. MAPLE also introduces Goal-Based Mission Planner for IMPACT to autonomously act upon.

4 Provenance Tracking

4.1 Dynamically Generated Provenance

Runtime generation of provenance typically requires applications to be instrumented accordingly [9]. Instrumenting applications and generating provenance at the same time can be cumbersome from a software engineering perspective [10]. To overcome this, we instrument the IMPACT system by treating it like a black-box and avoid including invasive code into the several services and bridges. All the services within IMPACT communicate using the ZeroMQ Distributed Messaging Framework to subscribe and process messages published by a data hub within IMPACT. These messages contain data about the scenario such as telemetry, vehicle task details, restricted zones, mission related events, etc. [6]. All messages being sent to and from the entire system goes through a data hub.

ProvPy is the Python implementation of the PROV-DM data model [11]. We use this toolkit to create a memory representation of provenance and serialize it [10]. ProvPy allows system provenance to be serialized into XML and JSON; it also provides capabilities to export the provenance to a graphical representation. The graphical representation of PROV-DM is a directed graph, which allows for common network metrics and analytics to be applied. We developed an IMPACT-Provenance service which subscribes to the data hub and listens to the entirety of the message traffic, generating provenance for messages of interest. The service dynamically converts IMPACT hub messages into the PROV-DM format.

4.2 Human Play Calling in IMPACT

An IMPACT operator calls a play by using a play workbook to accomplish different missions and tasks. The Fusion agent generates a set of solutions to potentially accomplish the tasks; by default, the most optimized solution is presented to be engaged, but the operator has the freedom to select other, less optimal solutions. The provenance service captures the interaction between the operator and the Fusion agent and generates provenance based on solutions and assets involved with the mission. Alternatively, an automatically generated play with pre-set parameters can be called from the TM to accomplish tasks at an even faster rate.

4.3 COMPACT Bridge Interactions

The COMPACT system communicates with IMPACT through a message bridge that acts as a service connected to the data hub. While air vehicles are accomplishing missions, COMPACT monitors their vehicle states and notifies IMPACT

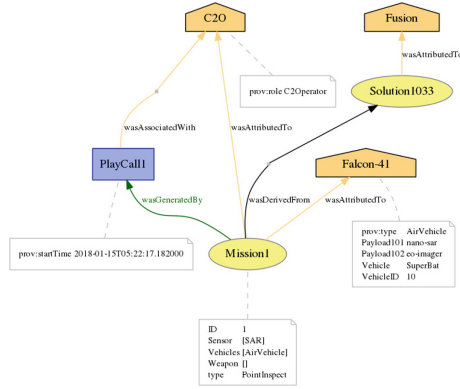


Fig. 4. Example of the Provenance of a Regular Play Call. The Play Call itself is an activity that generates the mission entity. The mission entity was derived from a set of solutions determined by the fusion agent and is attributed to the C2 operator and Falcon-41 unmanned air vehicle.

of policy violations when a vehicle leaves the communications range or enters a newly discovered restricted zone. Automated plays are then called with the collaboration of the PM and TM. When a vehicle is about to enter a restricted fly zone, the PM receives the warning from COMPACT (Fig. 4). Fusion will create a new restricted zone within IMPACT and attempt to reroute the vehicle. When a vehicle has violated a communications range policy, the PM receives the warning from COMPACT and requests a communications relay play, which the TM then executes (Fig. 5).

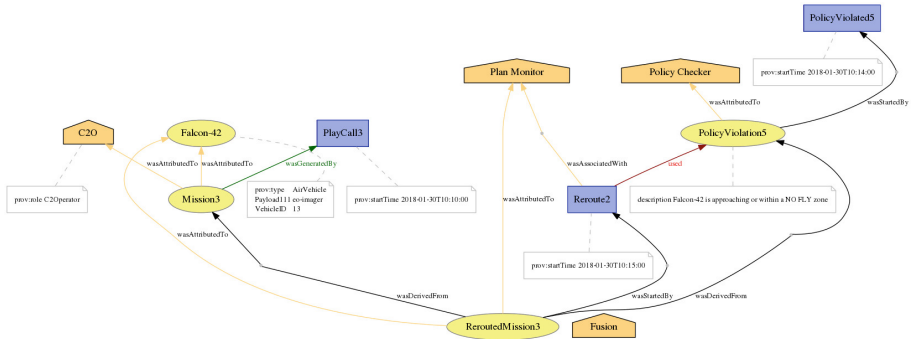


Fig. 5. Example of rerouted provenance. The policy checker agent represents COMPACT and detects that Falcon-42 is approaching a NO FLY zone and generates a policy violation. Plan monitor consequently does a rerouting activity for the operator.

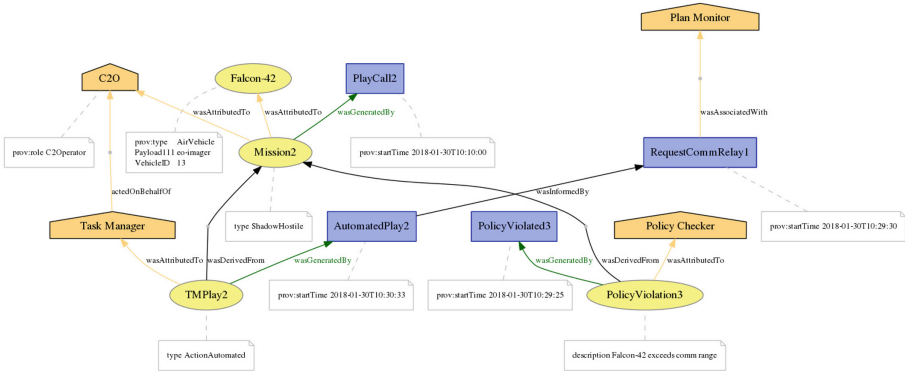


Fig. 6. Example of communications Relay Provenance. The Policy Checker (COM-PACT) detects that Falcon-42 is about to fly outside of communications range and generates a policy violation. Plan monitor then requests a communications relay play and task manager responds by creating a play to provide communications support.

4.4 MAPLE Bridge Interactions

Although the MAPLE framework allows IMPACT to communicate with live assets, provenance tracking cares more about the goal based missions it provides (Fig. 6). Some plays are enacted because of an overall mission being accomplished, such as base protection. In a scenario where an operator receives critical information regarding the location of hostile forces, the goal is to send assets to the location for reconnaissance. If that information turns out to be false, the asset needs to be dismissed from the mission in order to be reassigned to a new mission or to follow other leads to determine the initial reason for the assumption of hostile forces in the area. Provenance creates the relationships between the data in order to detect when plays have been invalidated and to notify the operator to free these assets. Furthermore, root cause analysis can be used on provenance data of a large scenario and can determine responsibility of false information or potential reasons as to why the information has been invalidated [12].

5 Transparency

Complexity increases dramatically in an ever-changing battlespace; situational awareness is difficult to maintain in these endeavors. Introducing provenance into adaptive systems is a critical step for system transparency and explanation. One of the key components for achieving algorithmic transparency is data provenance [13].

The data model provides a natural way to determine how data has evolved and been manipulated, which agents were responsible for different actions, and the processes involved. Systems should provide explanations regarding both the procedures followed by autonomy and the specific decisions that are made. We

Provenance can track the actions of an autonomous agent to provide operators with data lineage and process flow in order to enhance decision making and situational awareness. Similarly, the actions of human operators are also tracked to either provide responsibility flow or root cause reasoning.

We believe that accountability, transparency, and thus trust can be enhanced through the proper employment of provenance tracking. Provenance introduces a potential data model to bring explanation and transparency through different analytics techniques. This paper discussed a very small scale use of provenance, but opens the idea of using different analytic techniques to interface humans and machines using the data model.

References

1. Moreau, L., Groth, P.: Provenance: An Introduction to PROV Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan and Claypool, San Rafael (2013)
2. Moreau, L., Missier, P. (eds.): PROV-DM: The PROV Data Model. W3C Recommendation (2013)
3. Sezavar Keshavarz, A., Huynh, T.D., Moreau, L.: Provenance for online decision making. In: Ludäscher, B., Plale, B. (eds.) IPAW 2014. LNCS, vol. 8628, pp. 44–55. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16462-5_4
4. Huynh, T.D., Ebden, M., Venanzi, M., et al.: Interpretation of crowdsourced activities using provenance network analysis. In: The First AAAI Conference on Human Computation and Crowdsourcing, pp. 78–85 (2013)
5. Putnam, C., Waters, J., Rodas, O.: A standard decision format using provenance. In: 17th IEEE International Symposium on Signal Processing and Information Technology (2017)
6. Coronado, B., Gustafson, E., Reeder, J., Lange, D.S.: Mixing formal methods, machine learning, and human interaction through an autonomies framework. In: Proceedings of the 2016 AAAI Fall Symposium Series (2016)
7. Cummings, M.L.: Operator interaction with centralized versus decentralized UAV architectures. In: Valavanis, K.P., Vachtsevanos, G.J. (eds.) Handbook of Unmanned Aerial Vehicles, pp. 977–992. Springer, Dordrecht (2015). https://doi.org/10.1007/978-90-481-9707-1_117
8. Gutzwiller, R.S., Lange, D.S., Reeder, J., Morris, R.L., Rodas, O.: Human-computer collaboration in adaptive supervisory control and function allocation of autonomous system teams. In: Shumaker, R., Lackey, S. (eds.) VAMR 2015. LNCS, vol. 9179, pp. 447–456. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21067-4_46
9. Frew, J., Metzger, D., Slaughter, P.: Automatic capture and reconstruction of computational provenance. *Concurr. Comput. Practice Exp.* **20**(5), 485–496 (2008)
10. Moreau, L., Batlajery, B., Huynh, T.D., et al.: A templating system to generate provenance. *IEEE Trans. Softw. Eng.* (2017)
11. ProvPy: A Python library for W3C Provenance Data Model. <https://pypi.python.org/pypi/prov>
12. Jenkinson, G., Carata, L., Balakrishnan, N., et al.: Applying provenance in APT monitoring and analysis: practical challenges for scalable, efficient, and trustworthy distributed provenance. In: 9th USENIX Workshop on the Theory and Practice of Provenance (2017)
13. ACM US Public Policy Council: Statement on Algorithmic Transparency and Accountability. USACM Press Releases, 12 January 2017, pp. 1–2 (2017)



Cyber Officer Profiles and Performance Factors

Ricardo G. Lugo^{1(✉)} and Stefan Sütterlin^{2,3}

¹ Inland Norway University of Applied Sciences, Lillehammer, Norway
Ricardo.lugo@inn.no

² Faculty for Health and Welfare Sciences, Østfold University College, Halden, Norway
Stefan.sutterlin@hiof.no

³ Division of Clinical Neuroscience, Oslo University Hospital, Oslo, Norway

Abstract. The complex and uncertain nature of the cyber domain poses challenges for military to define and assess performance attributes needed for cyber operators. We propose using a cognitive engineering and human factors approach to develop proxies for performance indicators for complex human behavior in military cyber contexts. The Hybrid Space conceptual framework was developed and a series of studies was conducted to profile and predict cyber operator behaviors and performance. This included both micro- (cognitive styles) and macro-cognitive approaches (team workload demands). Results from these studies were then incorporated in cyber officer training through an OLB approach.

Keywords: Cyber · Human factors · Performance · Military

1 Cyber Domain in Military Operations

While cyber incidents raise increased media attention, the role of the human factor in cyber defense still lacks a comprehensive scientific framework [1]. Interdisciplinary approaches to investigate human factor in cyber defense operations include human interaction, the physical and social operating environment, decision-making processes and psychological determinants of performance in cyber officers [2, 3].

Rapid technological developments and definition of the cyber domain as a battlefield changed the cognitive demand profiles for cyber defense officers and challenged the traditional military organizational structures. Decisions on tactical levels can have large geo-strategical implications and are often highly complex, based on insufficient or unreliable information, and under time pressure.

The changed cognitive demands on cyber officers can be characterized by the requirement of enhanced cognitive agility, i.e., a parallel processing and constant monitoring of events, decisions and consequences in and between the intertwined physical and the cyber domain and their tactical and strategic dimensions. To assess, describe, and visualize the cognitive landscape in which officers operate, the Hybrid Space (HS) framework was developed [2]. The HS allows to assess performance in the cross-section of socio-technical and cyber-physical systems. This model was further expanded and tested by including metacognition, inter-individual differences, and macrocognitive (see Fiore et al. [33] for review) factors on HS performance.

1.1 Human Factors and Cyber Operations

Due to the lacking research within cyber defense, an applied paradigm to address this shortcoming is needed and from previous research, cognitive engineering is a human centered approach that addresses continuous changing complex environments [4]. This approach is supported through ecological approaches where participants are immersed in their domain, and the interaction of the person in their domain becomes the focus, unlike traditional experimentation. Such an approach supports the development of an understanding of influences of performance through an interactionist model. This is due to the cyber domain's high information load without any objective goal or measure [5]. It is important to understand the situational factors influence on the mental workload demands involved as these new situations (the cyber domain) in itself have no value, that is no correct choice, but a person's choice is dependent on the situational understanding and its influence on mental workload [6].

The goal of cognitive engineering is to develop an understanding of the fit of how an operator can better perform by taking more critical and complex decisions, in the system he finds himself in [5].

1.2 Cognitive Factors in Performance

The cognitive demands required for a successful cyber defense include a cognitive skill set with emphasis on cognitive flexibility, situational awareness, sustained attentional control, and motivation [7–9]. Being able to control one's attention is dependent on the ability to maintain alertness, being able to orient oneself to relevant sensory input, and to make decisions based on the perceived information [10].

But factors such as anxiety can inhibit attentional control on specific tasks thus reducing processing efficiency [11]. This is caused through inhibition dysfunction, where stimuli unrelated to a task able to enter attention, and by affecting attentional shifting, where irrelevant information requires conscious processing thus diminishing the efficiency of working memory. Emotion regulation describes the process of how one is able to emotionally respond to an ongoing situational that allows for flexible decision-making strategies [12]. This process begins when an emotionally relevant situation is encountered that requires attention to and appraisal of relevant events that require a response that is dependent on previous experience as well as current psychological and physiological states. This process occurs through antecedent strategies and response modulation. For a well trained cyber defense operator, this process could begin situational selection where they approach a threat and try to identify the source through gathering information from relevant sources. Due to the difficulty of predicting emotion responses to this threatening situation, the situation could elicit emotional responses from the operator that can lead to less desirable outcomes. This would lead operators to then modify the current situation by including emotional processing strategies. This then influences their attentional deployment, where one can either focus on emotionally relevant cues (i.e. increased heart rate) that can lead to preservative cognitions (rumination, worry) that often produce incorrect decision making behaviors. But if an operator is properly trained, being able to identify a maladaptive emotional response and

reappraising the situation, this leads to a reduction in physiological responses and subjective emotional responses, by reappraising the situation and thus leading to a more objective based decision making approach by understanding the impact of tactical decisions and strategic goals. If the decision taken leads to a negative outcome, for example, the network is breached and one cannot identify the source or stop the breach, an operator will try to modulate the emotional impact that they experience from the situation. This can be seen through increased sympathetic activation and a reduction in adaptive behaviors such as decreased communication and cooperation.

Metacognition refers to ‘thinking about thinking’ and includes the knowledge of one’s abilities, situational awareness, and behavioral regulation strategies [13]. Cadets with high metacognitive skills have more accurate and confident in their judgments performance within different situations. Individuals who have higher metacognitive awareness are also more accurate when describing their capabilities and identifying strategies that can improve performance. High metacognitive awareness of one’s cognitive processes involves monitoring, planning and evaluating one’s behavior in a given situation. If an operator is able to recognize the emotional impact of a breach and understands how the emotional response (e.g. rumination) affects their performance, they are better able to implement emotion-regulation strategies that would help to reappraise the situation and thus implement more objective decision making strategies.

Metacognition and emotion regulation can be viewed and intertwining processes that functions at a conscious level (awareness) that helps self-regulation, but also at a non-conscious level and co-regulates cognitions. Recent models [14, 15] of metacognition identify three distinct aspects of metacognition that also includes emotion regulation strategies that are based on declarative knowledge but also on subjective experiences based in emotional processes (intuition). These three aspects, metacognitive knowledge, experience and skills, work on three levels, social personal-awareness, and nonconscious level. While the social and personal-awareness level work through monitoring and controlling behavior (conscious level), the nonconscious level deals with emotion and cognition regulation.

1.3 Situational Factors on Performance

Due to the integration of cyber operations at each level in the military, from soldier support on the battlefield to network defense at higher command levels, cyber defense operators are dependent on team functioning. Their environment consists of several operators working in teams within a traditional military framework that responds to military leadership. Since this domain is novel to most personnel cyber operators must be able to work efficiently in these hybrid domains through proper communication and cooperation with both cyber and other military personnel. Research has identified several aspects of team functioning that can influence performance. Gutzwiller et al. [16] identified three areas that are crucial for performance: having an understanding and awareness of the network, which includes technical aspects and the behavior of the network; the world – how the physical world is affected or may be affected by events in the world including emergent threats and abnormal activities and behaviors; and finally the team, where awareness of work (completed, in-progress tasks), processes (demands, needs),

and bootstrapping, being able to communicate with other inter-agencies to maintain focus. Cooke et al. [17] proposes the Interactive Team Cognition (ITC) that cognition is an activity that can be studied at the team level but only in its contexts. This has led Salas [18] to identify several factors that influence team functioning. To increase team performance, training, and interventions directed at improving both cognitive and affective cognitions, teamwork processes and performance outcomes helps, interventions suited at improving team processes have better effects. More specifically, improving communication and cooperation skills lead to better outcomes than task training [19]. While communication and cooperation demands within teams helps performance, increased team support or team emotional demand have been shown to decrease performance. So if a situation is able to initiate and increase emotional content and unclarity within a team, team processes focused at emotional communication and team member support will be more dominant than task communication and cooperation and thus lead to worse performance.

2 Method

The presented series of studies applies cognitive engineering and psychological approaches and attempts to provide a better understanding of determinants and limitations of cyber operator performance and discusses its implications for selection, training, and testing.

Both micro- and macrocognitive approaches were used to profile cyber officer qualities and to understand how cognitive and situational factors would influence performance. Whereas microcognitive approaches can be seen as more basic research approaches, macrocognitive approaches are applied in nature. Researchers stress that both approaches are important and need to be included in naturalistic decision making paradigms to better capture cognitive processes, both individual and situationally induced [6, 20].

2.1 Microcognition

Microcognitive approaches have the most used approaches to experimental settings. Cognitive functions are usually assessed in artificial laboratory settings under controlled situations [21]. It has the advantage of isolating individual cognitive functions and allowing for identifying outcomes of specific manipulations. Historically, micro-cognitive approaches have helped identify processes in daily functioning to decision-making strategies. Microcognitive approaches are also important in uncovering traits that might be specific to the cyber domain. For this research, microcognitive approaches are used to identify leadership constellations, decision-making strategies, and how these influence performance. For profiling of personality traits, the Big Five Inventory [22], a self-constructed self-efficacy scale, Embedded figures test [23] and trait affect measurements [24] were used. Participants were also subjected to experimental paradigms such as emotional go-nogo [25] and a modified Cognitive Reflection Test [26]. For emotion regulation strategies and nonconscious metacognition, both psychophysiological

markers [27] (interoception) and cognitive aspects (rumination: response styles questionnaire [28], and the Penn state Worry Questionnaire [29]) were used. For the other metacognitive aspects, the Self-regulation Questionnaire [30] and the Metacognitive Awareness Inventory [31] was used to measure the conscious aspects.

But these approaches are limited in that they do not take into account external situational aspects that are found in natural environments that could influence behavior [20, 32].

2.2 Macrocognition

Macrocognition provides a framework to study cognitive processes as they affect real-world task performance, and is addressed as a complement, rather than a competitor, to microcognition by incorporating both individual and team processes [33] and is defined as “the internalized and externalized high level mental processes employed by teams to create new knowledge during complex, one of a kind, collaborative problem solving.” The term macrocognition emerged from a need to address the broad variety of cognitive processes in a naturalistic settings [20, 32, 34] and has gained recent focus through naturalistic decision making studies in sociotechnical systems [20, 33, 34]. Macrocognitive approaches are divided into three broad areas [21]: (1) macrocognitive modelling, where expert behaviors are compared to modelled systems, (2) macro cognition architectures, application of microcognitive functions to real world situations and, (3) team cognition, processes that arise due to several operators inter-acting in sociotechnical systems. To gather macrocognitive data, the Team Workload Questionnaire (TWLQ) was used [19]. The TWLQ was used to assess the workload demand in team tasks from two sub-scales: Team Workload Demands, which is concerned with the demands of team interactions (communication, coordination, team performance monitoring); and the Task-Team Workload Demands, which assesses the management of task and team workload demands (time share demands, team support demands, team emotion demands).

Measurements for dependent variables for naturalistic decision making paradigms (macrocognitive approaches) were developed. The Hybrid Space is mapped on a Cartesian plane and cyber operators marked their position simultaneously every hour during the third day of exercise (see Jøsok et al. [2] for description). In addition, students noted their current task at each position, to give context to further analysis. Movement in the Hybrid Space is operationalized through four constructs:

1. HSDT: distance traveled in the Cartesian Plane measured by Euclidian distance
2. HSxM: Movement along the cyber-physical domain (x-axis)
3. HSyM: Movement along the strategic-tactical domain (y-axis)
4. HSQC: Number of quadrant changes.

2.3 Data Collection and Analysis

Data for all studies was collected through self-reports and from a cyber-defense exercise where officer cadets were tested on breach detection and intrusion of a secure network.

Data was collected from mixed methods experimentation that included self-reports, expert evaluations, and naturalistic observation. Profiles were computed through

psychophysiological measurements (parasympathetic activation, interoceptive accuracy) and from self-reports from personality and cognitive styles (e.g. emotion regulation, field dependence/independence, self-efficacy, metacognition) inventories. Experts in cadet training were asked to rate officer candidates for leadership abilities and these were matched to personality profiles (Five Factor Model) and emotion regulation strategies (emotional response inhibition task). For macrocognitive outcomes for predictions in decision-making and performance in cyber defense, data was collected during the Norwegian Defense Cyber Academy's (NDCA) annual Cyber Defense Exercise (CDX): Exercise Cold Matrix. This arena facilitates the opportunity for students to train in tactics, techniques and procedures for handling various types of cyberattacks. The students work in teams, take tactical decisions in response to network intrusions, and develop counter measures. Success is presented as direct feedback to the decisions and actions taken during the exercise. Intrusions are initiated by an affiliated agency who are engaged to help the NDCA with their educational program.

3 Results

Lugo et al. [35–37] showed that cyber operators have different cognitive processes when compared to controls. They did not show the same associations on perseverative cognitions (rumination, worry) or emotion regulation processes. When tested for learning styles through the Embedded figures test, cyber cadet officers displayed distinct cognitive learning styles (field independent) versus age matched controls, but were similar to other non-cyber engineering students. Within the military domain, they were also significantly different than non-cyber military personnel. Participants who had higher self-efficacy responded incorrectly in decision-making strategies on more cognitive tasks when gut-feelings were involved.

External military experts rated cyber operators on leadership skills through situational observation and this was matched to personality profiles and emotion regulation strategies [38]. While previous findings show that military leadership reflects that of transformational leadership (higher extraversion, agreeableness, and openness, lower neuroticism) [39, 40], cyber operators showed that emotion regulation strategies moderated the relationship between extraversion and leadership ratings, but introverts were better rated [41]. These results partially reflect the findings of Rubin et al. [40] but theoretical aspects of why introversion better predicts leadership for cyber needs further investigation.

Metacognition was associated with better performance in cyber domain contexts [42]. Metacognition could predict overall movements in the HS except for performance on the strategic-tactical axis (y-axis). Strategic and tactical decisions are reliant on more macrocognitive approaches such as communication and cooperation between operators. Lugo et al. [38] used this to further investigate HS movements and showed that the team workload demands (communication, coordination) helped cyber operators with greater movements, while task-team workload demands (dissatisfaction, timeshare demands) inhibited movement.

4 Discussion

An understanding of cognitive, metacognitive and macrocognitive factors are needed to help develop effective and efficient cyber operators who find themselves in this domain. Current military structures may inhibit performance and therefore officers need to understand both their actions and positions, be able to identify significant others in the decision making process, and give the people responsible for decision-making the proper information that is found in the cyber domain so that decisions can be done with better certainty. Due to the fact that cyber operations understanding and research is in its infancy, the presented studies attempt to systematize an approach through the development of a conceptual framework, followed by the implementation of an action research approach by first identifying cognitive processes, and then incorporating situational variables, to give a better understanding of the qualities a cyber-operator may need to possess to successfully be able to perform in an complex and uncertain cyber domain.

Profiling of the officer cadets lead to several novel findings. Cyber engineer cadets do not display the same patterns of perseverative cognitions as normal controls do. The results from this study show that cyber domain officer cadets may differ in their rumination patterns from other comparable age groups. While normal controls showed signs of negative association between interceptive accuracy and perseverative cognitions (rumination, worry), cyber cadets did not show any of these associations. Cyber cadets did show similar cognitive styles (field independence) to that of other engineering students, but were significantly different from matched age controls and other military personnel (e.g. bomb diffusing units).

Due to the nature of the cyber domain being more oriented to more objective situations instead of containing emotional valence as found in physical domains, it was also expected, and found, that cadets who displayed higher confidence but included emotionally driven intuitions, performed worse than cadets who did not include intuition in cognitive tasks. While gut feelings can help in social situations where one must consider emotionally loaded content, the cyber domain might require more objective cognitions over emotional intuition, which could hinder the decision making process.

The cyber domain may lead to a selection process that attracts different profiles of cognitive and emotional processing. The relevance of individual differences in leadership constellations and cognitive styles may be from potentially systematic but unintended biases resulting from self-selection. Selection procedures are important to understand due to their implications for later job performance.

One important finding that arose from the investigations is the role of metacognition. Metacognitive awareness and regulation predicted performance. This is comparable to recent finding in similar domains in expert development. Metacognition is not an inborn talent, but rather a skill that is developed through training, exposure, and feedback and is a skill that distinguishes experts from novices.

Using a cognitive engineering and human factors approach, these results have led Knox et al. [43] to suggest the OLB model (“orient-locate-bridge”) for communication in a hybrid cyber physical domain, where cyber cadets are trained in technical aspects as well as psychological concepts of performance, are able to orient themselves, locate others, and bridge knowledge gaps that may interfere with decision-making. Thus

helping cyber officers understand the processes and influences of their behaviors, and be able to communicate and coordinate with others to improve the decision-making process. The OLB model incorporates human factors in the cyber domain to better understand the interaction of behaviors in the cyber environment, in a scientific approach that answers the need proposed by the scientific community [1, 3, 20, 44].

All of the studies used had limitations. The studies were correlational in nature due to the lack of conceptual frameworks and validated models. The approaches used in these studies were taken from other domains, e.g. team performance, and applied psychology such as education and clinical psychology and thus might not be applicable to this domain. Further research including these approaches is necessary. Dependent variables for the naturalistic decision making paradigm were constructed for these investigations.

5 Conclusion

The proposed chapter provides an overview of a series of comprehensive empirical research on determinants of cyber defense officer performance. Cyber operators show unique profiles, and their performance can be explained by both micro and macro approaches. In summary, these findings aim to pave the way for an evidence-based approach to selection, training and evaluation of cyber defense officer performance. To achieve this, the theoretical framework of the Hybrid Space has been applied to map the cognitive location and dynamics during cyber defense operations, leadership styles, problem-solving styles and group effects complement the overview of performance within the theoretical framework.

References

1. Gutzwiller, R.S., Fugate, S., Sawyer, B.D., Hancock, P.A.: The human factors of cyber network defense. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 59, no. 1, pp. 322–326. SAGE Publications, Los Angeles, September 2015
2. Jøsok, Ø., Knox, B.J., Helkala, K., Lugo, R.G., Sütterlin, S., Ward, P.: Exploring the hybrid space. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) AC 2016. LNCS (LNAI), vol. 9744, pp. 178–188. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39952-2_18
3. Mancuso, V.F., Christensen, J.C., Cowley, J., Finomore, V., Gonzalez, C., Knott, B.: Human factors in cyber warfare II emerging perspectives. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting (2014)
4. Woods, D.D., Roth, E.M.: Cognitive engineering: human problem solving with tools. *Hum. Factors* **30**(4), 415–430 (1988)
5. Gersh, J.R., McKneely, J.A., Remington, R.W.: Cognitive engineering: understanding human interaction with complex systems. *Johns Hopkins APL Tech. Dig.* **26**(4), 377–382 (2005)
6. 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)
7. Helkala, K., Knox, B., Jøsok, Ø.: How the application of coping strategies can empower learning. In: 2015 IEEE Frontiers in Education Conference (FIE), pp. 1–8. IEEE, October 2015

8. Helkala, K., Knox, B., Jøsok, Ø., Knox, S., Lund, M.: Factors to affect improvement in cyber officer performance. *Inf. Comput. Secur.* **24**(2), 152–163 (2016)
9. Helkala, K., Knox, B., Jøsok, Ø., Lugo, R., Sütterlin, S.: How coping strategies influence cyber task performance in the hybrid space. In: Stephanidis, C. (ed.) *HCI 2016. CCIS*, vol. 617, pp. 192–196. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40548-3_32
10. Posner, M.I., Petersen, S.E.: The attention system of the human brain. *Annu. Rev. Neurosci.* **13**(1), 25–42 (1990)
11. Eysenck, M.W., Derakshan, N., Santos, R., Calvo, M.G.: Anxiety and cognitive performance: attentional control theory. *Emotion* **7**(2), 336 (2007)
12. Gross, J.J.: The emerging field of emotion regulation: an integrative review. *Rev. Gen. Psychol.* **2**(3), 271 (1998)
13. Jacobs, J.E., Paris, S.G.: Children’s metacognition about reading: issues in definition, measurement, and instruction. *Educ. Psychol.* **22**(3–4), 255–278 (1987)
14. Efklides, A.: Metacognition: defining its facets and levels of functioning in relation to self-regulation and co-regulation. *Eur. Psychol.* **13**(4), 277 (2008)
15. Efklides, A.: Interactions of metacognition with motivation and affect in self-regulated learning: the MASRL model. *Educ. Psychol.* **46**(1), 6–25 (2011)
16. Gutzwiller, R.S., Hunt, S.M., Lange, D.S.: A task analysis toward characterizing cyber-cognitive situation awareness (CCSA) in cyber defense analysts. In: 2016 IEEE International Multi-disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), pp. 14–20. IEEE, March 2016
17. Cooke, N.J., Gorman, J.C., Myers, C.W., Duran, J.L.: Interactive team cognition. *Cogn. Sci.* **37**(2), 255–285 (2013)
18. Salas, E., DiazGranados, D., Klein, C., Burke, C.S., Stagl, K.C., Goodwin, G.F., Halpin, S.M.: Does team training improve team performance? A meta-analysis. *Hum. Factors* **50**(6), 903–933 (2008)
19. Sellers, J., Helton, W.S., Näswall, K., Funke, G.J., Knott, B.A.: Development of the team workload questionnaire (TWLQ). In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1, pp. 989–993. SAGE Publications, Los Angeles, September 2014
20. Klein, G., Moon, B., Hoffman, R.R.: Making sense of sensemaking 2: a macrocognitive model. *IEEE Intell. Syst.* **21**(5), 88–92 (2006)
21. Smieszek, H., Rußwinkel, N.: Micro-cognition and macro-cognition: trying to bridge the gap. In: *Proceedings of the 10th Berlin Workshop on Human-Machine Systems: Foundations and Applications of Human-Machine Interaction*, pp. 335–341 (2013)
22. Soto, C.J., John, O.P., Gosling, S.D., Potter, J.: Age differences in personality traits from 10 to 65: big five domains and facets in a large cross-sectional sample. *J. Pers. Soc. Psychol.* **100**(2), 330 (2011)
23. Witkin, H.A.: *A Manual for the Embedded Figures Tests*. Consulting Psychologists Press, Palo Alto (1971)
24. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* **54**(6), 1063 (1988)
25. Hare, T.A., Tottenham, N., Galvan, A., Voss, H.U., Glover, G.H., Casey, B.J.: Biological substrates of emotional reactivity and regulation in adolescence during an emotional go-nogo task. *Biol. Psychiatry* **63**(10), 927–934 (2008)
26. Frederick, S.: Cognitive reflection and decision making. *J. Econ. Perspect.* **19**(4), 25–42 (2005)
27. Schandry, R.: Heart beat perception and emotional experience. *Psychophysiology* **18**(4), 483–488 (1981)

28. Treynor, W., Gonzalez, R., Nolen-Hoeksema, S.: Rumination reconsidered: a psychometric analysis. *Cogn. Ther. Res.* **27**(3), 247–259 (2003)
29. Meyer, T.J., Miller, M.L., Metzger, R.L., Borkovec, T.D.: Development and validation of the penn state worry questionnaire. *Behav. Res. Ther.* **28**(6), 487–495 (1990)
30. Carey, K.B., Neal, D.J., Collins, S.E.: A psychometric analysis of the self-regulation questionnaire. *Addict. Behav.* **29**(2), 253–260 (2004)
31. Schraw, G., Dennison, R.S.: Assessing metacognitive awareness. *Contemp. Educ. Psychol.* **19**(4), 460–475 (1994)
32. Klein, G., Ross, K.G., Moon, B.M., Klein, D.E., Hoffman, R.R., Hollnagel, E.: Macrocognition. *IEEE Intell. Syst.* **18**(3), 81–85 (2003)
33. Fiore, S.M., Rosen, M.A., Smith-Jentsch, K.A., Salas, E., Letsky, M., Warner, N.: Toward an understanding of macrocognition in teams: predicting processes in complex collaborative contexts. *Hum. Factors* **52**(2), 203–224 (2010)
34. Letsky, M., Warner, N., Fiore, S.M., Rosen, M., Salas, E.: Macrocognition in complex team problem solving. Office of Naval Research Arlington, VA, June 2007
35. Lugo, R., Sütterlin, S., Helkala, K., Knox, B., Jøsok, Ø., Lande, N.M.: Interoceptive sensitivity as a proxy for emotional intensity and its relation to perseverative cognition. *Psychol. Res. Behav. Manag.* **11**, 1–8 (2017)
36. Lugo, R.G., Sütterlin, S., Knox, B.J., Jøsok, Ø., Helkala, K., Lande, N.M.: The moderating influence of self-efficacy on interoceptive ability and counterintuitive decision making in officer cadets. *J. Mil. Stud.* **7**(1), 44–52 (2017)
37. Lugo, R., Iversen, Ø., Sütterlin, S.: Learning styles contribution to decision-making in cyber defense officers. Manuscript under preparation (2017)
38. Lugo, R., Sütterlin, S., Helkala, K., Knox, B., Jøsok, Ø., Lande, N.M.: The relationship between personality and leadership in cyber defence cadets. Manuscript under preparation (2017)
39. Lim, B.C., Ployhart, R.E.: Transformational leadership: relations to the five-factor model and team performance in typical and maximum contexts. *J. Appl. Psychol.* **89**(4), 610 (2004)
40. Rubin, R.S., Munz, D.C., Bommer, W.H.: Leading from within: the effects of emotion recognition and personality on transformational leadership behavior. *Acad. Manag. J.* **48**(5), 845–858 (2005)
41. Lugo, R., Kwei-Nahr, P., Jøsok, Ø., Knox, B.J., Helkala, K., Sütterlin, S.: Team workload demands influence on cyber detection performance. Manuscript under preparation (2017)
42. Knox, B.J., Lugo, R.G., Jøsok, Ø., Helkala, K., Sütterlin, S.: Towards a cognitive agility index: the role of metacognition in human computer interaction. In: Stephanidis, C. (ed.) *HCI 2017. CCIS*, vol. 713, pp. 330–338. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58750-9_46
43. Knox, B., Jøsok, Ø., Helkala, K., Khooshabeh, P., Ødegaard, T., Kwei-Nahr, P., Lugo, R.G., Sütterlin, S.: Socio-technical communication: the hybrid space and the OLB-model for science-based cyber education. Manuscript accepted for publication *Military Psychology* (2017)
44. Tikk-Ringas, E., Kerttunen, M., Christopher, S.: Cyber security as a field of military education and study. *Jt. Force Q.* **75**, 57–60 (2014)



Displaced Interactions in Human-Automation Relationships: Transparency over Time

Christopher A. Miller^(✉)

Smart Information Flow Technologies (SIFT),
319 First Avenue N., Suite 400, Minneapolis, MN 55401, USA
cmiller@sift.net

Abstract. Transparency (roughly, the provision of information about what the automated system is doing and why, potentially at multiple levels of abstraction and goal directedness) in automated systems has repeatedly been shown to improve human-machine interaction and performance, as well as human acceptance and trust. Nevertheless, there is a fundamental problem with providing transparency information at the time of action execution: specifically, that excessive human workload, which typically motivates the inclusion of automated systems, may not permit the absorption of the transparent information. We propose “displacing” the provision of transparent information in time and/or space from the time of execution and show how this approach is tied to beneficial findings for pre-mission planning and post-mission debriefing and explanations in human-automation interaction.

Keywords: Transparency · Explanation · Debriefing · Mission planning
Team interactions · Trust · Cognitive workload

1 Introduction

As the automated systems with which humans interact become more and more “autonomous” they, largely by definition, operate at times and in contexts in which human supervisor/controllers are not actively monitoring or issuing control inputs. This may be because inputs are impossible (such as in loss of communications due to jamming or weather or deep space communication time lags) or because humans are not able, willing or expected to intervene (e.g., due to workload constraints, inferior performance or simply to expectations of autonomous functioning).

These circumstances represent a separation of human input from execution, rather than a complete removal. Humans still design and build the systems and set them up (perhaps in a “mission planning” phase) to operate “autonomously” within certain constraints and for certain objectives. In the end, we humans still want the systems to behave for our benefit and within the instructions we provide.

Yet this displacement of control means that behavior shaping inputs and interactions will, increasingly, be displaced relative to action execution. The displacement may well be geographical—after all, one of the uses for unmanned systems is to have them operate in locations where it is dangerous for humans to go—but it will *inevitably* be temporal. We will be tasking automation increasingly via abstract guidelines, plans,

and policies well before actions are taken and decisions are made, and we will be reviewing those actions and decisions (and, ideally, explanations for them) after they have been completed and their effects are known.

This means that what we have known about human-automation relationships must be re-examined for “displaced interactions”. We must ask questions such as the effect of varying degrees of temporal and geographic displacement on maintaining Situation Awareness (and what that means when communications are not possible or expected), what skills or personal traits are needed to accurately, effectively and safely interact with automated systems over increasing degrees of displacement, and what kinds of decision aids and training may make such interactions easier and more effective.

We have begun thinking about one such “displacement effect”—the role of transparency in displaced interactions. “Transparency”, here, refers to the ability for the automation to be inspectable or viewable in the sense that its mechanisms and rationale can be readily known. Researchers since at least Billings [1] and Sarter and Woods [2, 3] have called for greater transparency in automation. Furthermore, recent research has generally confirmed [4–7] that such transparency yields better human situation awareness, trust and, frequently, better overall human-machine performance than systems which include less or no transparency.

But there is a problem inherent in transparency that begs for an answer to the displacement question. As noted previously [8] not all information about what an “autonomous” agent is doing and why can be shared if there is to be any workload savings in a multi-agent team. Indeed, for the human to attend to and process any information about what an “autonomous” agent is doing will come at the expense of the human’s attention devoted to perceiving and understanding other aspects of their world and performing actions within it. We should expect this to produce additional workload and potential loss of situation awareness of other aspects of the work context if attention is oversubscribed.

Since automation is frequently introduced precisely to enable a human operator to do more within available time, workload (and training), the introduction of additional information for the human to process during this busy period may be particularly problematic. Worse, as noted above, there are contexts such as deep space exploration and military operations under communications jamming in which the communication of information to support automation transparency may simply not be possible.

So, we can conclude that transparency is valuable in human-automation interaction, but information to support it frequently cannot be communicated—at least in the moment when it is most needed—for a variety of reasons. Is there a way out of this dilemma?

2 Displaced Transparency

The problem represented by this dilemma is hardly new, nor unique to human-automation interactions. Humans interacting with other humans have encountered and wrestled with it throughout time. Human supervisors attempting to increase their capabilities through organizing and administering (human) subordinates have almost identical requirements: the need to maintain awareness of the performance of

their subordinates so as to ensure that the supervisor’s intent is accomplished as accurately as possible—even when that subordinate must necessarily act at a geographic or temporal distance from the supervisor.

As I have argued elsewhere [9], acting through a human subordinate is a process of delegation. Delegation inherently involves the communication of intent with oversight (including inspection and correction) of the subsequent performance of that intent by the subordinate. Furthermore, the communication of intent itself forms an “intentional frame” which can serve to increase situation awareness and decrease communication and cognitive processing demands in the future. Sheridan’s original definition of supervisory control [10] included the provision that supervisors had to communicate their intent to subordinates or, in Sheridan’s words, to “teach” subordinate automation what it should do.

Human to human communication in and near the moment of execution is an extraordinarily rich tool, especially when it is deployed against the backdrop of common cultural and professional understanding of the domain and its goals and methods, and even moreso when it is augmented by mutually-understood professional jargon. It serves to make the communication of intent from supervisor to subordinate, and status from subordinate to supervisor, extraordinarily efficient and rich. But it is worth noting that there is evidence that high-performing human-human teams, especially in high criticality domains, frequently exhibit less (and less explicit) communication than do less well-integrated teams [11].

Part of the reason human natural language communication can be so effective, and a large part of the reason well-trained and experienced teams can be so sparse with their communications, is that such team members share an understanding of the domain and of work within it. This understanding is certainly acquired during training and experience, but on a day by day (or mission by mission) basis, it is also acquired through mechanisms such as training and planning before execution, and explanation and debriefing (or after-action reviews) after execution. Below, I will argue that these mechanisms provide “transparent” information to the human team-members—that is, the same kind of information that has been shown to improve performance from “transparent” displays. They just do so at times other than the time of execution. That is, they provide temporally (and potentially, geographically) displaced information about how an agent is, will, or should behave, or how and why he/she/it did behave, but do so at a time when workload and attentional demands are lower (potentially on both supervisor and subordinate) than the time and context of execution. In short, they provide *Displaced Transparency*.

3 Transparency and Why Displaced Information Can Provide It

Chen et al. [12] have defined transparency as “...the descriptive quality of an interface pertaining to its abilities to afford an operator’s comprehension about an intelligent agent’s intent, performance, future plans, and reasoning process.” But the emphasis on “the interface” in the above definition puts a, perhaps undue, focus on information that is conveyed during execution—when an interface is typically used. I contend that much

information which achieves the goals of transparency (i.e., affording “an operator’s comprehension about an intelligent agent’s intent, performance, future plans and reasoning process”) need not be provided only, or even primarily, at that workload-intensive time.

Indeed, Chen [12] also defines a scale or model for transparency, the Situation Awareness-based Transparency (SAT) scale, which in turn leverages Endsley’s [13] scale for Situation Awareness. Chen’s SAT levels are summarized below as transparency information intended to support:

1. Level 1 SA (What’s going on and what the agent is trying to achieve) which is satisfied by providing information about the agent’s:
 - Purpose or Desire and Goal selection
 - Process and Intentions (including Planning and Execution) and Progress along that process
 - Performance of both the process and in general.
2. Level 2 SA (Why does the agent do what it does?) which is satisfied by providing information about the agent’s:
 - Reasoning process for planning or decision making, including the agent’s beliefs and broader purpose, including agent’s current beliefs about the environmental and other factors which constrain it
3. Level 3 SA (What should the operator expect to happen?) which is satisfied by information about the agent’s:
 - Projection to Future/End state
 - Potential Limitations including likelihood of error and history of performance.

If we take this information content as what is required for effective transparency, then it is worth noting that much of it could be—and in much effective human-human teaming, is—provided either before or after the time of action execution. “What an agent is trying to achieve?” is something that can and generally should be worked out, at least at a high level, before the subordinate agent is deployed. Intent expressions (which may include goals, purposes, methods and priorities [14]), by their nature, occur before action, while an understanding of why an agent does what it does and what its beliefs were that might have influenced decisions and actions are precisely the focus of the explanations that occur in effective after-action reviews and debriefings [15]. Figure 1 provides a hypothesized annotation of Chen’s SAT levels into items which can occur before or after the moment of action.

Figure 2 provides a simple timeline for a pre-planned workflow (e.g., for a military reconnaissance mission) to illustrate the point. If we assume mission execution to begin at time T, let us say that at some point before T (i.e., time T minus n), a supervisor and subordinate plan this recon mission as consisting of an Ingress Phase to begin at time T and end at time T+1:00 h, to be followed by a Search Phase to run from the end of the Ingress Phase for another hour. These will then be followed by a decision point whose agreed-upon logic is that if a target has been detected it will be Monitored for another hour and if not, the agent will Traverse to an egress point.

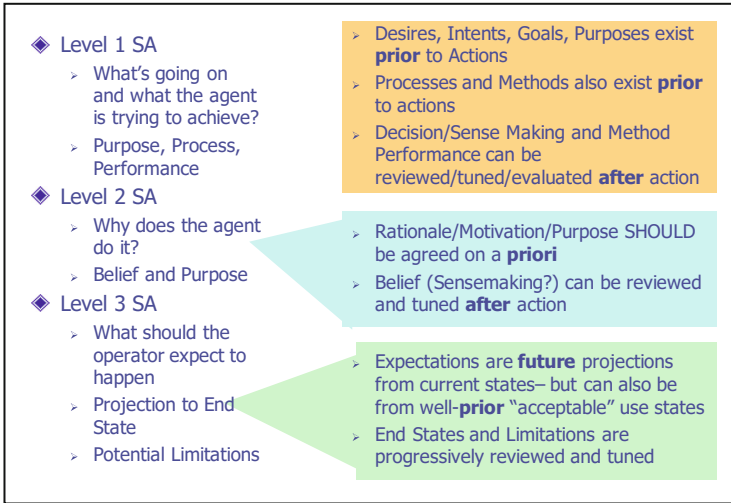


Fig. 1. Chen et al.'s [12] SAT levels annotated for their ability to be conveyed either before or after the time of action execution.

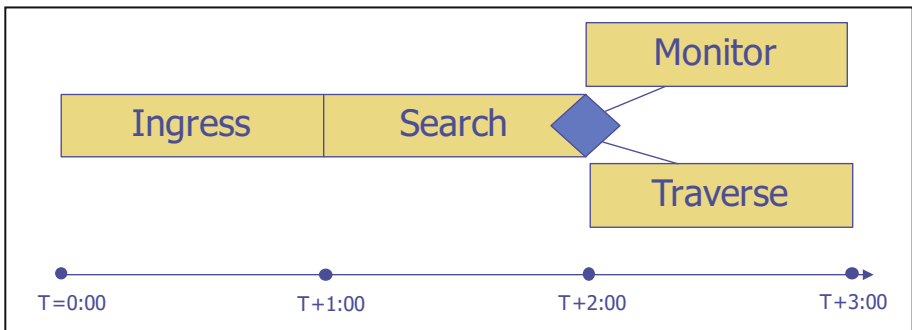


Fig. 2. Simple timeline example to illustrate conveyance of “transparent” information before use via planning.

The simple fact of having made this a prior plan affords substantial situation awareness to the supervisor *even if s/he has no further communication* with the subordinate. For example:

- Given that it is 45 min into the mission, the supervisor knows that the subordinate is engaged in Ingress and even, approximately, where the subordinate is. This is “What’s going on?” knowledge—Level 1 SAT.
- Furthermore, the supervisor knows why the subordinate is ingressing: to get to the search area and begin search—Level 2 SAT.
- Finally, the supervisor knows that, at time 1:00, the subordinate will transition to Search. This is “What the operator should expect to happen?” knowledge—Level 3 SAT.

Granted, all of this SA can be in error if no further communication is available. Even the best laid plans can go awry, and lack of communication will make that deviation opaque (the opposite of transparent) to the supervisor. But we should also note that having an a priori plan in place makes communication more efficient. Instead of having to report all three levels of SA (what is going on, why, and what to expect next), the subordinate will generally have to report only current status or, perhaps, only deviations. The presence of pre-planned alternatives and acceptable error bounds (as represented partially by the pre-planned decision point in Fig. 2) makes it possible to further reduce communication needs. If, anywhere from time 1:00 to 2:00 the subordinate reports finding a target, then the behavior at the time 2:00 decision point becomes predictable. Even more, if at some time after 2:00 the subordinate reports that it is Monitoring, the superior is entitled to conclude that the reason is because a target was found during Search. Even in the event where the entire plan is made impracticable, having had a shared plan makes communication about future behaviors more efficient by giving all participants a shared ground to build from. If, for example, a fuel leak makes sustained Monitoring impossible, both supervisor and subordinate will know (at least approximately) where the agent is, how much range is required to get home, that a target has been detected, that the monitoring objective will have to be abandoned, etc. And all of this knowledge is shared with little or no “in the moment” information exchange or processing.

The above example emphasizes the role of pre-mission planning in establishing “displaced transparency”, but that is not the only way transparent information can be displaced. After action reviews, debriefings and explanations also provide after-the-fact transparent information as well. Admittedly, this information is not provided in a timely fashion to enable a supervisor to override or correct behavior which may not be desired, but insofar as work with this particular subordinate continues in the future, it does play a mutual learning and trust building/tuning role. By learning how the subordinate thought and behaved in a specific situation, and potentially by offering advice or instruction about how future instances should be handled, the supervisor can shape future behavior in much the same way that planning shapes behavior (and information interpretation) before execution. As Tannenbaum and Cerasoli [15] have said, after-action debriefs are an effective and efficient means of improving team performance and team cohesion. Their meta-analysis of debriefing studies (covering 46 samples and 2,136 individual participants) indicated that on average, debriefs improve effectiveness over a control group by approximately 25% ($d = .67$).

Figure 3 illustrates the “displaced transparency” relationship we posit. Most transparency research to date focuses on information which is conveyed in a narrow temporal window around an action or event of interest. Such research has generally shown improvements in performance, situation awareness and trust when transparency information is provided. But such research has rarely examined workload effects, especially in time critical and overloaded periods and/or has examined it with subjective and coarse-grained tools such as NASA TLX [16]. We posit that there are periods in human-automation interaction, just as there are in human-human interaction, where the human’s attention, processing and comprehension capabilities are so sparse and/or over-subscribed, that the attention to transparent information will, at best, be incomplete, and at worst may provide a disastrous distraction.

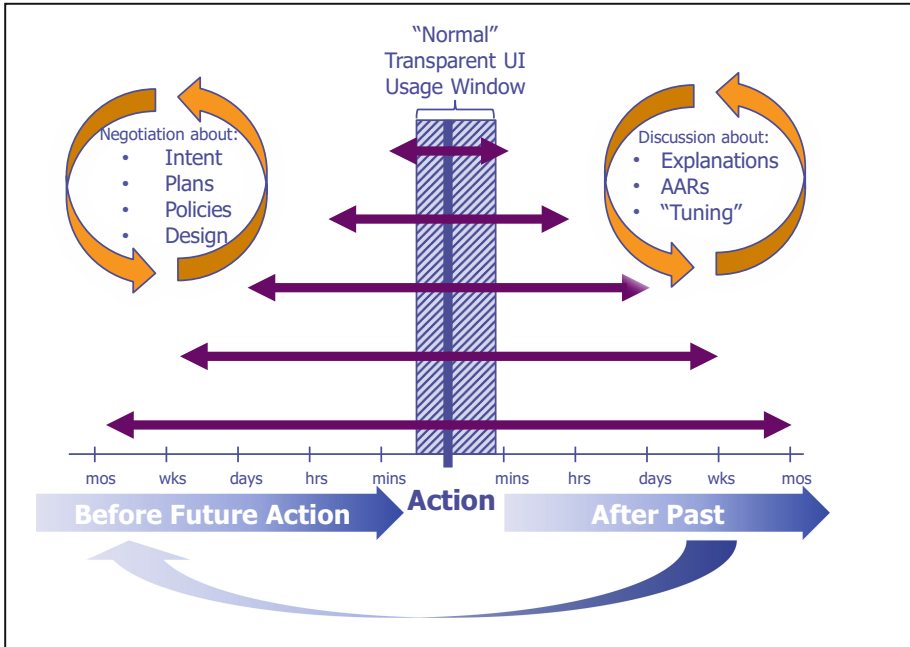


Fig. 3. Conceptual relationship between “in-action” transparent information and “displaced” transparent information.

We suspect, however, that it will be possible to displace much of this transparent information in time into periods which are substantially less overloaded. These periods may be anywhere from months or years before an action, or weeks to months after an action. The processes by which this transposition occurs are called training, discussion, planning, etc. when they occur before an action and the communication is about intent, plans, policies, alternatives and how to decide between them. At extreme durations, the “planning” process becomes one of design of the team, its concept of operations or its equipment. When the discussion occurs after the action, it is called explanation, debriefing, after-action review or various forms of “tuning” (including continuous quality improvement). Iterative cycles through either pre- or post-action discussion will likely only help the understanding of transparent information among team members. It is worth noting, in addition, that the process of pre-action planning and post-action review itself forms a virtuous cycle that can build team understanding for future actions. Although one might reasonably ask how and why providing transparent information *after* an action is helpful, the answer is that it will be helpful in providing team coordination and understanding for future actions, assuming that team must interact again in the future.

4 A Mechanism for Displaced Transparency and Its Effects

The claim that displacing transparent information can serve much the same purpose as presenting it at the time of need begs the question of how such a process might work. The work of Kambhampati, Chakraborti, Talamadupula and others [17, 18] may provide at least the beginnings of an answer.

Although working to provide explanations from machine or robot planners to humans, they nevertheless begin with a human social and cognitive model of the role of explanation itself. They take issue with many past approaches to providing machine explanations as rooted in presenting the machine's reasoning in its own terms—a process they call “soliloquy.” [17] They say “Such soliloquy is wholly inadequate in most realistic scenarios where the humans have domain and task models that differ significantly from that used by the AI system.” Instead, they say, effective explanation between actors must be a “Model Reconciliation Problem” [17].

Kambhampati has pointed out [19] that except perhaps in testing circumstances, explanations are generally neither sought nor provided in circumstances where individuals believe that their reasoning and behaviors are compatible or “reconciled”. In other words, if I believe that I understand your mental model of the situation sufficiently that you and I would arrive at the same decision about a course of action to pursue, then no explanation of that decision is necessary. It is only when our models diverge (or are believed to diverge) that explanations are invoked—either because you do something which is not compatible with my model of what should be done in the situation, or because you intend to do something that you believe will differ from my model in the situation.

Explanation is thus, in their view, largely a process of Mental Model Reconciliation. It is provided in order to synchronize the models of those who must work together. It is sought when models do not synch—and it may not be sought when models are assumed to synch, even if they do not.

While explanations can certainly deviate from our actual methods of decision making [20, 21], they nevertheless represent how we are trained and acculturated to providing rationalizations for our decision making. Thus, it represents a form of team or group synchronization of thinking in its own right.

This viewpoint explains many phenomena in both explanation and transparency. For example, the fact that well-performing, efficient teams require less, not more, explicit communication [11] can be seen as arising from the fact that such teams are likely to have trained and worked together extensively in the past. Thus, their mental models of task and domain are likely to be well-synchronized—meaning that explanatory interactions are less likely to be required, commands can be abbreviated with contextual modifications presumed, and even task-based status information can be abbreviated and given with reduced contextual information because all parties are likely to understand what is needed when.

Similarly, team debriefing after a mission or project, with most of its associated benefits [15] can be attributed to the fact that effective debriefings involve team members interacting about their decisions and their behavioral processes. Simply knowing who knew what when and how they made decisions on the basis of that knowledge serves to educate other team members about their teammates' mental models, while group discussion about how things might be done more effectively in the future creates shared mental models going forward. Group discussion and shared rationalization of events and processes can also serve an educative function, effectively causing the group to converge

on a series of broad thought processes and values that, within the “culture” formed by the group, count as valid and reasonable “ways of thinking” (or, at least, of explaining) [22]. Note, though, that this can cut both ways—resulting in adverse cases in “group-think” [23] and “normalization of deviance” [24]. Finally, explanations offered within a group or to superiors play a social function as well [25], serving to reduce social distance and establish or reinforce power structures. While these functions may not directly contribute to a shared mental model about how decisions should be made within the group, they will serve to enhance team cohesion and affiliation.

On the pre-mission planning side, Dwight Eisenhower famously said “Plans are worthless, but planning is everything” [26]. One sense in which this is undoubtedly true is supported by the notion of model reconciliation. Extensively pouring over plans, rehearsing what could go wrong and what might be needed in contingent situations is a process which conveys and affords opportunities for model synchronization in much the same way that post-mission debriefing does. Participants are likely to emerge from such a process with a richer understanding of each other’s models—and having had many opportunities to confront and refine those models in a process which tends toward synchronization. While this, too, can engender groupthink and the suppression of some voices, it also allows team members to “get inside each other’s heads” and, thereby, increase their chances of knowing how each other will behave even in unanticipated situations.

Note that I am, in no way, claiming that measuring and computing model reconciliation needs and effects will be easy, especially between humans and machines. Chakraborti and Kambhampati and their colleagues have largely sidestepped this problem by using machine readable symbolic models for “simulated” humans in order to illustrate efficiency gains in symbolically characterized explanation content. By contrast, humans have evolved cultural, semantic and pragmatic markers rooted in natural language and “body language” for interpreting the need for model reconciliation and then effecting it—a process that, though complex and rich, is far from error proof (cf., [27] for examples of human-human and human-machine model mismatch where communication failed to avoid misinterpretations and, therefore, accidents).

5 Example of Efficiencies Through Model Reconciliation

It is reasonably straightforward to show how a process of model reconciliation which occurs either before or after a time frame in which the model is used to make a decision can lead to a reduction in the need to provide “transparency” information in that time frame. Consider again the simple mission sketched in Fig. 2, along with an “Observing Teammate” (OT) who, let’s assume, can observe and know everything that is happening in the mission context but is completely unaware of the mission and the intentions of an “Enacting Teammate” (ET). In this scenario, any reported intentional interpretation of behavior from the ET will certainly increase OT’s SA, but that is because OT’s knowledge is essentially nil without such reports. Even observable events in the world (e.g., a headwind) will need to be interpreted for OT in order to provide Level 1 SA (Purpose, Process, Performance) knowledge, since the ability to interpret world events for their impact on the plan will be non-existent. OT has none of ET’s mental model of the mission (though they may share a model of the world state in

this example). All three levels of Chen's SAT [4] must be communicated for OT's full SA. On the other hand, if OT and ET share the same mental model and can perceive the same world events, then ET might need to communicate nothing since events would be interpreted similarly and decision making would be identical between ET and OT.

Even when OT can't observe everything that ET can (e.g., in the case of a remote supervisor), the burden of communication is substantially reduced. When events are unfolding as planned, at most ET might need to communicate confirmations that the plan is proceeding as expected. Even when unexpected events occur (e.g., the headwind above), reporting them may be all that is necessary to synchronize ET and OT's models of the impact and revisions necessary and desirable to the mission (revisions in knowledge at Level 3 SA that will *not* have to be communicated since both sides will make them concurrently—though confirmation might still represent useful redundancy). As before, anticipated variations (e.g., the decision point about whether to remain and monitor or simply to traverse) can be communicated much more tersely since both parties know, a priori, what the significance of a detected target will be on this decision point, and/or what the valid reasons are for remaining to monitor vs. traversing.

Even mental model mismatches become easier to detect given this prior planning. Let's say that OT failed to notice that ET detected a target and thus, doesn't understand why ET is transitioning to Monitor rather than Traverse. Simply posing the question in the context of what was expected to be a shared model of the mission identifies the mismatch. "Why are you Monitoring?" conveys a violation of expectations for Monitoring (i.e., the prior detection of a target) and hones in on the piece of information which is needed to repair model mismatch.

Finally, although harder to quantify, post-mission debriefings which are later followed by subsequent missions can have similar effects. If, for example, OT learns that ET has a tendency toward speedy completion of missions, s/he might assume a bias or preference in ET for Traversing vs. Monitoring and, in the presence of an ambiguous target detection signal, make more nearly accurate predictions about what ET will do. This represents a variation in the mental model (specifically, in values or priorities) between ET and OT, but insofar as OT understands this about ET (that is, OT's model of ET contains it) it will be accurately factored in to OT's SA and result in accurate understanding of the situation.

6 Predictions and Next Steps for Displaced Transparency

Above, I have presented an argument for the effects and desirability of displacing the presentation of transparency information into a priori mission planning interactions and a posteriori explanations and debriefings, along with a hypothesized mechanism for why these effects might be obtained. We know, from the sources cited above and others, that transparency information frequently provides detectable mission performance benefits. We also know (again from sources sited) that prior mission planning and explanation and debriefings also provide benefits for team cohesion, team satisfaction and team performance. It seems likely that these benefits obtain because they are making use of the same underlying mechanism: the communication of information which promotes situation awareness through mental model synchronization at the time

of use. The fact that this information doesn't have to all be transmitted at the time of use, but instead can be spread into lower workload periods before and after usage, is a feature we should use more extensively in design—for human-human and also for human-machine interactions.

Some simple, testable predictions from this model are provided below. While we have not yet been able to conduct experiments to validate these predictions, a simple laboratory test seems eminently plausible. A relevant yet simple scenario is sketched in [18] where a human and a robot are located in a building with a long corridor and multiple side rooms. The human tasks the robot to fetch “a med kit”—one or more of which may be located in the side room(s). Which med kit is desired, expected, and provided is a function of elements of context (e.g., where robot, human, med kit(s) and other potential humans using med kits are located) as well as the robot's decision making algorithm. Of course, mismatches in elements of mental models (e.g., awareness of the physical context and of the robot's decision making process) are exactly what is required to provide SAT knowledge—and can be manipulated in an experimental design. The human may expect the robot to go to an different side room if s/he erroneously believes the med kit to be located there, or the robot may take longer and travel further than necessary if it has an erroneous model of where the human will be located. In this or a similar paradigm, we would predict:

- With mental model synchronization between teammates, reduced time, workload effort and even communication bandwidth will be necessary to achieve a similar level of situation awareness compared to conditions without mental model synchronization.
- Shifting the communication of transparent information into other time frames (before or after execution) will yield improved situation awareness (with reduced workload) even under conditions of communications restriction or constrained workload for the human recipient, given that model synchronization makes that information comprehensible.
- A priori mental model reconciliation will produce more accurate inferences by team members even in unanticipated situations, even with little or no explicit communication of transparent information.
- Particularly with regard to post-mission debriefing and explanations, effects of a posteriori model reconciliation will produce increased awareness and ability to predict teammates behavior even in unanticipated situations in subsequent missions.

We note with interest that [18] reports that the inclusion of mental modeling capabilities in the reasoning of a robot agent, where the robot was modeling the expected reasoning of a human operator and reacting accordingly, produced a 44–75% improvement in robot decision making in terms of avoiding resulting resource conflicts in one analytic experiment they performed.

What is less well documented is the tradeoffs involved in shifting transparency information into time frames before and after it is needed. Somewhere between “no plan survives first contact with the enemy” (implying that “overplanning” is wasteful) and “Plans are worthless but planning is everything” (implying that planning activities are very valuable), there must lie a (probably context-dependent) happy medium. Where is that medium, and what parameters characterize it? It is likely that information

theoretic models can provide us with boundary conditions for this claim, but their relationship to actual human-human (or human-machine) interaction remains to be determined.

Finally, as automation becomes more capable, omnipresent and more “autonomous” in complex work domains, it is becoming clearer that it cannot provide all sufficient transparency information *in the moment* of action execution. Even if the human is capable of understanding it given time, it will all too frequently be the case that s/he will be engaged in other tasks and will be unable to devote sufficient attention and cognitive processing capability in a timely fashion. Instead, we need to strive to enable automation to participate in pre-mission planning and in post-mission debriefing and explanations in order to develop and accurately tune human trust and comprehension frameworks so that available capacity in the moment of use will be sufficient.

Acknowledgments. I am indebted to Dr. Jesse Chen for providing a forum for initial thoughts on this topic, and to Rao Kambhampati for the insight that explanations generally need to focus only on mismatches in mental models.

References

1. Billings, C.: *Aviation Automation: The Search for a Human-Centered Approach*. Erlbaum, Mahwah (1997)
2. 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)
3. Sarter, N.B., Woods, D.D., Billings, C.E.: Automation surprises. *Handb. Hum. Factors Ergon.* **2**, 1926–1943 (1997)
4. Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D., Procci, K.: Intelligent agent transparency in human-agent teaming for Multi-UxV management. *Hum. Factors* **58**(3), 401–415 (2016)
5. Lyons, J.B., Havig, P.R.: Transparency in a human-machine context: approaches for fostering shared awareness/intent. In: Shumaker, R., Lackey, S. (eds.) *VAMR 2014. LNCS*, vol. 8525, pp. 181–190. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07458-0_18
6. Osofsky, S., Sanders, T., Jentsch, F., Hancock, P., Chen, J.Y.: Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In: *SPIE Defense + Security*, p. 90840E. ISOP (2014)
7. de Visser, E.J., Cohen, M., Freedy, A., Parasuraman, R.: A design methodology for trust cue calibration in cognitive agents. In: Shumaker, R., Lackey, S. (eds.) *VAMR 2014. LNCS*, vol. 8525, pp. 251–262. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07458-0_24
8. Miller, C.A.: Delegation and transparency: coordinating interactions so information exchange is no surprise. In: Shumaker, R., Lackey, S. (eds.) *VAMR 2014. LNCS*, vol. 8525, pp. 191–202. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07458-0_19
9. Miller, C., Parasuraman, R.: Designing for flexible interaction between humans and automation. *Hum. Factors* **49**(1), 57–75 (2007)
10. Sheridan, T.: Supervisory control. In: Salvendy, G. (ed.) *Handbook of Human Factors*, pp. 1244–1268. Wiley, New York (1987)
11. Entin, E., Serfaty, D.: Adaptive team coordination. *Hum. Factors* **41**, 312–325 (1999)

12. Chen, J.Y., Procci, K., Boyce, M., Wright, J., Garcia, A., Barnes, M.: Situation awareness-based agent transparency (No. ARL-TR-6905). ARL/HRED Aberdeen Proving Ground, MD (2014)
13. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. *Hum. Factors* **37**(1), 32–64 (1995)
14. Miller, C.: Delegation for single pilot operation. In: 2014 HCI-Aero. ACM, New York (2014)
15. Tannenbaum, S.I., Cerasoli, C.P.: Do team and individual debriefs enhance performance? A meta-analysis. *Hum. Factors* **55**(1), 231–245 (2013)
16. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: *Advances in Psychology*, vol. 52, pp. 139–183. North-Holland (1988)
17. Chakraborti, T., Sreedharan, S., Zhang, Y., Kambhampati, S.: Plan explanations as model reconciliation: moving beyond explanation as soliloquy. In: *Proceedings of IJCAI*, pp. 156–163 (2017)
18. Talamadupula, K., Briggs, G., Chakraborti, T., Scheutz, M., Kambhampati, S.: Coordination in human-robot teams using mental modeling and plan recognition. In: *IROS 2014*, pp. 2957–2962. IEEE (2014)
19. Kambhampati, S.: Personal communication, Arlington, VA, 1 August 2017
20. Tversky, A., Kahneman, D.: Judgment under uncertainty: heuristics and biases. *Science* **185** (4157), 1124–1131 (1974)
21. Klein, G.: Naturalistic decision making. *Hum. Factors* **50**(3), 456–460 (2008)
22. Miller, C.: Learning to disagree: argumentative reasoning skill in development. Ph.D. thesis. University of Chicago, August 1991
23. Turner, M.E., Pratkanis, A.R.: Twenty-five years of groupthink theory and research: lessons from the evaluation of a theory. *Organ. Behav. Hum. Decis. Process.* **73**(2–3), 105–115 (1998)
24. Vaughan, D.: *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press, Chicago (1997)
25. Brown, P., Levinson, S.C.: *Politeness: Some Universals in Language Usage*, vol. 4. Cambridge University Press, Cambridge (1987)
26. Eisenhower, D.D.: From a Speech to the National Defense Executive Reserve Conference in Washington, D.C., 14 November 1957. Eisenhower, D.D.: *Public Papers of the Presidents of the United States*, p. 818. National Archives and Records Service, Government Printing Office (1957). https://en.wikiquote.org/wiki/Dwight_D._Eisenhower. Accessed 3 Mar 2018
27. Miller, C.: Social relationships and etiquette with technical systems. In: Withworth, B., de Moor, A. (eds.) *Handbook of Research on Socio-Technical Design and Social Networking Systems*, Information Science Reference, Hershey, PA, pp. 472–486 (2009)



Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance

Mark A. Neerincx^{1,2(✉)}, Jasper van der Waa¹, Frank Kaptein²,
and Jurriaan van Diggelen¹

¹ TNO, Kampweg 55, 3769 DE Soesterberg, Netherlands

{mark.neerincx, jasper.vanderwaa, jurriaan.vandiggelen}@tno.nl

² Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, Netherlands
f.c.a.kaptein@tudelft.nl

Abstract. Most explainable AI (XAI) research projects focus on well-delineated topics, such as interpretability of machine learning outcomes, knowledge sharing in a multi-agent system or human trust in agent’s performance. For the development of explanations in human-agent teams, a more integrative approach is needed. This paper proposes a perceptual-cognitive explanation (PeCoX) framework for the development of explanations that address both the perceptual and cognitive foundations of an agent’s behavior, distinguishing between explanation generation, communication and reception. It is a generic framework (i.e., the core is domain-agnostic and the perceptual layer is model-agnostic), and being developed and tested in the domains of transport, health-care and defense. The perceptual level entails the provision of an Intuitive Confidence Measure and the identification of the “foil” in a contrastive explanation. The cognitive level entails the selection of the beliefs, goals and emotions for explanations. Ontology Design Patterns are being constructed for the reasoning and communication, whereas Interaction Design Patterns are being constructed for the shaping of the multimodal communication. First results show (1) positive effects on human’s understanding of the perceptual and cognitive foundation of agent’s behavior, and (2) the need for harmonizing the explanations to the context and human’s information processing capabilities.

Keywords: Explainable AI · Human-agent teamwork · Cognitive engineering Ontologies · Design patterns

1 Introduction

Advances in Artificial Intelligence (AI) and Information & Communication Technology (ICT) have been manifested in various automated systems, such as sensing technology, machine learning modules, Internet of Things, conversational agents and cognitive robotics. The embodiment in artificial, virtual or physical, agents enables automation to evolve as a member of mixed human-agent teams. A major challenge is to combine, automate and embody the information processes in such a way that agents really become full-fledged team-members, complementing and collaborating with the human team-members.

The coordination and collaboration in human-agent teamwork requires intelligent reactive and anticipatory behaviors of the agents. More specifically, they require a shared

knowledge representation, methods to comply with policies and agreements for responsible teamwork, and the learning and effectuation of successful patterns of joint activities (e.g., [1, 11, 18, 25]). In addition, according to the fifth challenge of Klein et al. [17], the agent should be able to make pertinent aspects of their status and intentions obvious to their teammates. In our view, this means that there is a need for mutual human-agent exchange of the reasons and foundations of actions, and that the agent should provide explanations for adequate human understanding and appreciation (incl. trust) of its performances.

Symbolic AI, such as BDI-agents with built-in Beliefs, Desires and Intentions based on folk psychology, provides explicit opportunities for the generation of explanations that are understandable and useful for the human team-member. For example, there have been developed explanation methods for fire-fighting teams [10], and tactical air combat teams [12, 28]. However, intelligent agents will be embodying more-and-more sub-symbolic machine learning methods for which it is far from clear how to derive an explanation logic for the human-agent collaboration.

So, for the envisioned human-agent teamwork, we need to develop methods for explainable AI (XAI) for adequate human understanding and appreciation (including trust) of symbolic *and* sub-symbolic agent performance. Such explanation should support human-in-the-loop (co-)learning of the human-agent teams.

To meet this challenge, our research focuses on the development of complementary explanation methods for agents using a holistic approach which covers all three phases of an explanation (see Fig. 1).

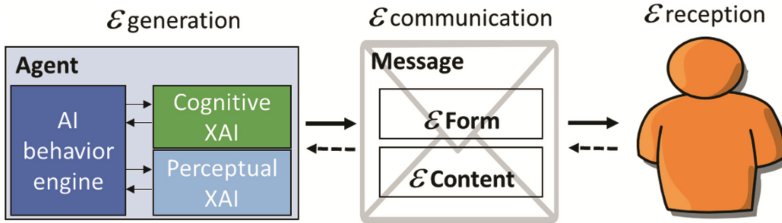


Fig. 1. Different phases of an explanation (ϵ refers to “the concept of explanation”, cf. [29])

The first phase concerns ϵ -*generation*. In this phase, we distinguish between Perceptual XAI and Cognitive XAI. Perceptual XAI aims to explain the perceptual foundation of the agent behavior, and is usually connected to the sub-symbolic reasoning parts of an AI architecture. This paper discusses two types of perceptual XAI: The construction of an Intuitive Confidence Measure and the classification of facts and foils for contrastive explanations [30–32]. The second type is cognitive XAI, which can explain why a certain action was chosen (e.g., by relating them to goals or beliefs; [13–15]). This part interacts with the sub-symbolic part of the AI behavior engine to ground its belief base. The construction of goal-based deliberative explanations is an example of cognitive XAI.

The second phase in the explanation process is ϵ -*communication*. This process is characterized by the form of explanation (e.g. textual, or using dedicated images as proposed by Beller et al. [2]), and the content of the explanation (i.e. what is

communicated). We will propose the use of ontologies to standardize the language which is used to provide explanations [27, 29]. For the design of the form of the explanations (which need to be adaptive and interactive), we are developing interaction design patterns (cf., [23, 33, 34]).

The third phase of explanation, *ϵ -reception*, concerns how well the human understands the explanation. With respect to XAI reception, some user studies (e.g., [22]) have been conducted, but there is a significant lack of empirical research with actual human task performers who need explanations in realistic human-agent settings (e.g., [20, 21]). This paper summarizes some first results of our evaluations.

We are developing a general, perceptual-cognitive, explanation framework that can be applied and refined in different projects, crossing domains. In this way, we can provide technological progress and empirical grounding, building general models and methods for explanation that can be instantiated in different domains. First prototypes have been developed and tested with end-users in transportation, healthcare and defense domains. Results show that human needs and preferences for explanations depend on their individual characteristics (e.g., age and experience) and on the operational context. The paper will present an overview of the explanation framework and the first prototype designs and evaluation outcomes.

2 Perceptual XAI

We propose two methods for perceptual XAI: (1) The construction of an Intuitive Confidence Measure (ICM) and (2) the identification of the counterfactual reference (i.e., the foil that is set against the fact in a contrastive explanation).

2.1 Intuitive Confidence Measure

Waa et al. [30, 31] developed a generic confidence or certainty measure for agent's machine learning (ML) that can be understood by the human. It is model-agnostic, i.e., the measure can be used for any machine learning model as it depends solely on the input and output of a trained model and future feedback about that output. The confidence (or uncertainty) reflects machine learning model's expected performance on a single decision or classification. Our Intuitive Confidence Measure (ICM) should be easily understood by humans without ML-knowledge, and behave in a predictable way. We designed ICM to be intuitive by basing it on the notion of similarity and previous experiences: Previous experiences with the ML model's performance directly influence the confidence of a new output, and this influence is based on how similar those past data points are to the new data point (similarity can be represented as a distance in an n -dimensional space). The ICM is applicable to any (semi-)supervised machine learning model, that is either trained online or offline, by treating it as a black box. One of the use cases is an agent that predicts if its remote teammate is required to be at location in the near future (e.g. at a work-desk in a dynamic positioning ship). The agent can explain how likely it is that this single prediction (output) is correct [31].

We have applied the ICM explanation method to the domain of monitoring dynamic positioning systems, i.e. highly automatic systems which aim to maintain a vessel's position and heading using dedicated propellers and thrusters [6]. Occasionally, the system requires human intervention, for example to warn the user about potential problems that are predicted by a machine learning model. Because these predictions may be wrong, the user must have an appropriate level of trust in this type of advice. We used the intuitive confidence measure to provide this kind of advice. The message below is an example of this type of ϵ -communication, where the ICM measure in the advice is communicated to the user. Furthermore, the particular design pattern behind the message also allows extra information to be provided on request of the user (in the example on the type of *changes*, and the type of *conditions*) (Fig. 2).

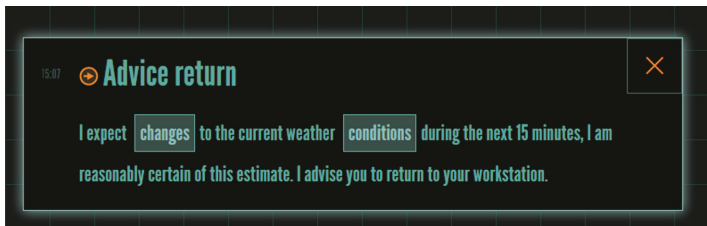


Fig. 2. Advice accompanied by the ICM XAI method.

To test the ϵ -reception of the ICM XAI method, we performed a user experiment [31], where we compared ICM with two different methods of computing confidence measures. The results confirmed the expectation that the intuitiveness of the XAI method is an important factor to consider when designing explainable smart systems. The user study showed large individual variations concerning user's interest in a confidence value and its utility in different situations. In comparison with other confidence measures, the ICM proves to be relatively easy to understand although not always preferred. The main reason for this was that the users, without ML-knowledge, attributed desired properties to other measures they did not fully understand. Whereas, with the ICM they understood the measure well enough to even identify its weaknesses; the users quickly thought to understand complex methods while, in fact, they did not (see [31], for more details of the user study).

2.2 Counterfactual Reference

For humans, explanations are most often contrastive, i.e., they refer to specific counterfactual cases, called foils [20]. When asking why an agent made a specific prediction or decision, humans would like to be informed about the contrast against the prediction or decision. So, an explanation should compare agent's output to an alternative counterfactual output (i.e., the foil). In other words, a good explanation considers both the fact (output) and the foil (alternative output).

Therefore, the generic PeCoX method for automated explanation extraction (i.e., ϵ -generation) includes the identification of the foil for a contrastive explanation. This

method should be able to deal with the context-dependency by learning which foil matches with which input-output pair based on interaction feedback from the user (e.g., through one-shot learning in combination with high generalization). This method is an extension of past work that constructs an explanation with the help of a localized interpretable machine learning model, such as a decision tree, through error weighting based on the distance between the data point of interest and the rest [24]. This way, the explanations can become more focused with less redundant information as the ‘why’ question is answered in a more precise manner.

More specifically, the PeCoX foil identification method is inspired by the LIME method, which provides an algorithm that learns an interpretable model locally around the prediction and a method to explain models by presenting representative individual predictions [24]. Further, like the Intuitive Confidence Measure of Sect. 2.1, LIME is model-agnostic. In our framework, it is further important to classify the foils for the generation of *adaptive* explanations (i.e., attuned to a specific user group).

The content and format of the explanations should be well-tailored to the user group. Ontology design patterns are constructed (see Sect. 4) that specify the high-level feature set, such that feature-based explanations are easy to comprehend by the user. Furthermore, interaction design patterns are generated that specify the corresponding multi-modal dialogues for contrastive explanations.

We are applying the contrastive explanation method to the domain of type1 diabetes mellitus (T1DM). T1DM is a chronic condition where insufficient insulin is produced by the pancreas, affecting blood glucose levels. Daily self-management is needed for a balanced glucose level, harmonizing insulin doses (via pen or pump), food intake (amount of carbohydrates) and activities (e.g. sport). High and low glucose levels may lead to a hypo or hyper, and on the long term to serious health complications. The *Personal Assistant for a healthy Lifestyle* (PAL¹) project develops a system for children aged 8–14, their parents and health care professionals that advances child’s diabetes self-management. The PAL system comprises an *Embodied Conversational Agent*

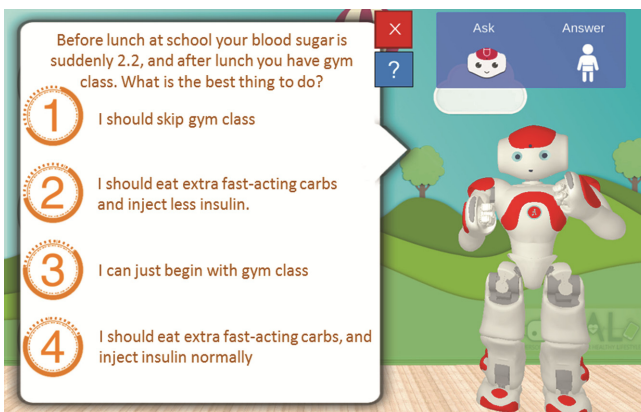


Fig. 3. Example of PAL quiz-question for learning to act on a low blood sugar level.

¹ <http://pal4u.eu>.

(ECA) robot and (mobile) avatar (Fig. 3), a set of *mobile health* (mHealth) applications (e.g., diabetes diary, educational quizzes), and *dashboards* for the caregivers (i.e., health care professionals and parents). All parts are interconnected with a shared knowledge-base and reasoning mechanism. health care professional, parent or child). Example contrastive explanations concerns PAL predictions that a child will not complete his or her self-management tasks. Such predictions should be explained to the child, parent and healthcare professional in different ways. Section 3 will provide more concrete explanation examples of the PAL system.

3 Cognitive XAI

In this part of the framework we consider explanations from the *intentional stance* [5]. When taking the intentional stance, you assume the action is a consequence of the intentions of the agent performing the action. You then explain the action by giving the reasons for the underlying intention. An explanation like such typically consists of beliefs, goals, and/or emotions [4, 5, 7, 9, 19]. For example, a support agent that tells its user to eat vegetables every day might provide the following explanation: ‘I hope (emotion) that you will take my advice to eat vegetables every day because I want (goal) you to adopt a healthy lifestyle, and I think (belief) that you currently do not eat enough vegetables’. A method for explaining the reasons is being developed for a cognitive-affective (BDI-based) agent that updates its beliefs, goals and emotions based on events perceived in the environment [13, 14]. This agent can explain and justify its actions by communicating (a) the beliefs that underpin the actions, (b) the goals that inform the human of the agent’s desired state when acting, and (c) the emotions that trigger or shape the actions.

3.1 Goal- and Belief-Based Explanations

The reasoning of a BDI based agent often consists of many beliefs and goals. If we use all of those for the explanation then this might overflow the user with information, which would thus not help us to make the behaviour intelligible [16]. Current work in XAI for artificial agents has thus focused on finding guidelines for which beliefs and goals are most valuable to use in an explanation towards end-users [3, 10]. This work confirms that a good explanation is short and thus contains few beliefs and goals. Similar to the evaluation of the Intuitive Confidence Measure in Sect. 2.1, a user study showed individual differences in explanation preferences. Particularly, adults showed a stronger preference than children for goals over beliefs in the explanations [14]. So, individual characteristics of the user must be taken into account when choosing which beliefs and goals should be selected as content for the explanation (Fig. 4).

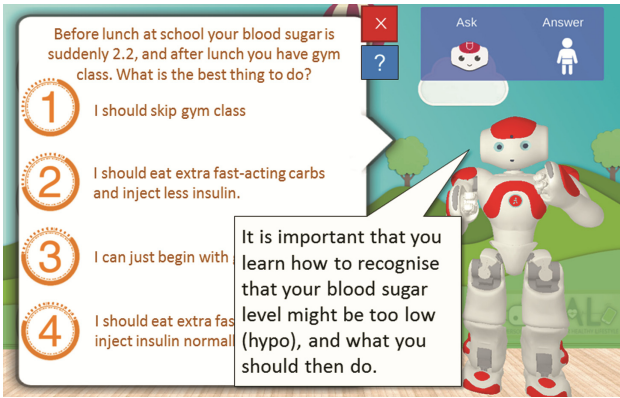


Fig. 4. Example of goal-based explanation in PAL.

3.2 Emotion-Based Explanations

Many artificial agents use computational modelling of emotion (based on cognitive appraisal theory) for their behaviour [15]. Here it is proposed that computational modelling of emotion can enhance XAI in several ways. The previous section mentioned the difficulty of selecting beliefs and goals as content for the explanation. The first use of emotions in XAI is as heuristic to identify the most important beliefs and goals for explanation. For example, the goal with the strongest influence for computing the desirability of an event can be used to explain the action the agent did after perceiving the event. The second use of emotions in XAI addresses that humans often use emotions when explaining behaviour [7]. The simulated emotions can be used to provide or enrich the action explanation (e.g., I was disappointed and therefore stopped my action). The third use addresses that emotions themselves can require to be explained (e.g., I was disappointed, because I had to do the same task over and over again) (Fig. 5).

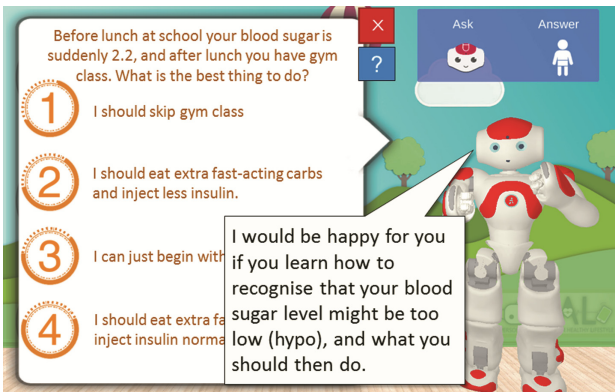


Fig. 5. Example of emotion-based explanation in PAL.

4 Ontology and Interaction Design Patterns

In our research and development projects, human-agent teams are modeled as joint cognitive systems, in which ontologies define the knowledge structures of these systems (i.e., the concepts, with their relationships, which underpin the cognitive processes of the teamwork). The explicit modeling of key concepts in explanations, enhances knowledge sharing (cf., [27]). These ontological models can be specified at different levels of generality [26]. The PeCoX top-level ontologies model general concepts that are domain independent, including explanation and human factors concepts (such as confidence and emotion, respectively). Specific domain and task aspects are captured in corresponding (lower level) ontological sub-models. For the construction of generic adaptive explanations, the Perceptual-Cognitive XAI framework is providing an extendable PeCoX-ontology in the form of *Ontology Design Patterns* (ODP, [8, 29]).

A formal definition of an explanation, like ODP, should specify its components, their roles and their interactions. Two ontologies provide starting points for such a definition. First, Su et al. [27] present an explanation ontology for constructing explanations for two agents that need to come to a partial shared understanding. They distinguish two knowledge types for agents that are engaged into an explanation process: (1) The *domain knowledge* entails representations of the to-be-explained concepts with the explanation parameters, and (2) the *explanation knowledge* entails the concepts that describe the explanation process and the permitted interactions in this process. Second, Tiddi et al. [29] studied approaches of Cognitive Sciences to model explanations with the instantiations in each of the analyzed disciplines. Their ontology intends to provide an abstract description which can be applicable to any context where an agent automatically produces explanations.

The ontologies of Su et al. [27] and Tiddi et al. [29] do insufficiently address (1) the perceptual XAI foundation of ϵ -generation, and (2) the situated human needs for explanation, i.e., the ϵ -reception (see Fig. 1). Section 2 describes the perceptual XAI concepts that are being specified and included in the PeCoX ontology. For the ϵ -reception, the social sciences provide theories and “models” for specific human information processes, for example, on the perception and activation of intentional and affective behaviors [20]. We are focusing and formalizing these “sub-models” into ODPs that support the predictions of ϵ -reception outcomes.

Whereas OPDs are used to support the design and reasoning of the communication processes and content of explanations, we use *Interaction Design Patterns* (IDP) for the specification of the form or shape of this communication [23, 33, 34].

5 Conclusions

This paper presented an integrative development approach for explainable AI in human-agent teams, addressing the perceptual and cognitive foundations of an agent’s behavior during the explanation generation, communication and reception. The proposed PeCoX framework is being developed and tested in the domains of transport, health-care and defense. This framework distinguishes between perceptual and cognitive explanations.

On a technological side it provides tools for generating explanations from perceptual components (often implemented using sub-symbolic or connectionist approaches, such as neural nets), and cognitive components (often implemented using symbolic approaches, such as rule-based and ontology-based knowledge systems). On a human-side it provides tools for understanding the different types of explanatory support that a human would want in different contexts on both a cognitive as well as a perceptual level. Ontology Design Patterns are being constructed for the reasoning and communication of explanations, whereas Interaction Design Patterns are being constructed for the shaping of the adaptive, multimodal communications.

First results of the PeCoX framework development were acquired in the domains of transport, health-care and defense. At the *perceptual level*, first, the Intuitive Confidence Measure (ICM) proves to be a good candidate to enhance human's understanding of the perceptual and cognitive foundation of agent's behavior in a dynamic position ship. The ICM-evaluation showed the need to attune the explanations to the context and the biases of human's perception of their own understanding. The participants in the evaluation of different confidence measures quickly thought to understand complex methods while, in fact, they did not. The explanation should be formulated in a "human-aware" way that minimizes the risk for such misbeliefs. Furthermore, the "foil" identification seems to be an effective method to generate desired contrastive explanations and situate them in a relevant context.

At the *cognitive level*, the goal, belief and emotion models provide the analytical knowledge foundation of the explanations. Similarly to the "ICM evaluations" at the perceptual level, the cognitive evaluations showed a need for further adaptation of explanations to the momentary context (e.g., age and role of the user). The user studies provide the empirical foundation of the explanation adaptation (i.e., the harmonization of belief-, goal- and emotion communication to the momentary user state and context).

It should be noted that, for effective and efficient explainable AI, a balance must be found between technological possibilities (which types of explanations can be provided by the underlying control logic of the autonomous system?), and user needs (given the current context, which type of explanations must be given for the human to establish an appropriate level of trust in the system?). This might involve making the technology better explainable, or accepting that the user cannot be explained in every aspect, and trying to mitigate possible negative consequences.

Acknowledgements. This research is supported by the European PAL project (Horizon2020 grant nr. 643783-RIA), and the TNO seed Early Research Program "Applied AI".

References

1. Bradshaw, J.M., et al.: From tools to teammates: joint activity in human-agent-robot teams. In: Kurosu, M. (ed.) 2009 Proceedings of the HCD 2009. LNCS, vol. 5619, pp. 935–944. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02806-9_107
2. Beller, J., Heesen, M., Vollrath, M.: Improving the driver–automation interaction: an approach using automation uncertainty. *Hum. Factors* **55**(6), 1130–1141 (2013)

3. Broekens, J., Harbers, M., Hindriks, K., van den Bosch, K., Jonker, C., Meyer, J.-J.: Do you get it? User-evaluated explainable BDI agents. In: Dix, J., Witteveen, C. (eds.) *MATES 2010. LNCS (LNAI)*, vol. 6251, pp. 28–39. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16178-0_5
4. Churchland, P.M.: Folk psychology and the explanation of human behavior. In: Greenwood, J. (ed.) *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge University Press, Cambridge (1991)
5. Dennett, D.C.: Three kinds of intentional psychology. In: Healey, R. (ed.) *Reduction, Time and Reality*. Cambridge University Press, Cambridge (1981)
6. van Diggelen, J., van den Broek, H., Schraagen, J.M., van der Waa, J.: An intelligent operator support system for dynamic positioning. In: Fechtelkötter, P., Legatt, M. (eds.) *AHFE 2017. AISC*, vol. 599, pp. 48–59. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60204-2_6
7. Döring, S.A.: Explaining action by emotion. *Philos. Q.* **53**, 214–230 (2003)
8. Gangemi, A., Presutti, V.: Ontology design patterns. In: Staab, S., Studer, R. (eds.) *Handbook on Ontologies. IHIS*, pp. 221–243. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_10
9. Harbers, M., Broekens, J., van den Bosch, K., Meyer, J.J.: Guidelines for developing explainable cognitive models. In: *Proceedings of ICCM*, pp. 85–90, January 2010
10. Harbers, M., van den Bosch, K., Meyer, J.J.: Design and evaluation of explainable BDI agents. In: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, vol. 2, pp. 125–132. IEEE (2010)
11. Hayes, B., Shah, J.A.: Improving robot controller transparency through autonomous policy explanation. In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 303–312. ACM (2017)
12. Haynes, S.R., Cohen, M.A., Ritter, F.E.: Designs for explaining intelligent agents. *Int. J. Hum Comput Stud.* **67**(1), 90–110 (2009)
13. Kaptein, F., Broekens, J., Hindriks, K.V., Neerincx, M.: CAAF: a cognitive affective agent programming framework. In: Traum, D., Swartout, W., Khooshabeh, P., Kopp, S., Scherer, S., Leuski, A. (eds.) *IVA 2016. LNCS (LNAI)*, vol. 10011, pp. 317–330. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47665-0_28
14. Kaptein, F., Broekens, D.J., Hindriks, K.V., Neerincx, M.A.: Personalised self-explanation by robots: the role of goals versus beliefs in robot-action explanation for children and adults. In: *RO-MAN 2017* (2017)
15. Kaptein, F., Broekens, D.J., Hindriks, K.V., Neerincx, M.A.: The role of emotion in self-explanation by cognitive agents. In: *DFAI Workshop at ACII 2017* (2017)
16. Keil, F.C.: Explanation and understanding. *Annu. Rev. Psychol.* **57**, 227–254 (2006)
17. Klein, G., Woods, D.D., Bradshaw, J.M., Hoffman, R.R., Feltovich, P.J.: Ten challenges for making automation a “team player” in joint human-agent activity. *IEEE Intell. Syst.* **19**(6), 91–95 (2004)
18. Lohani, M., Stokes, C., Dashan, N., McCoy, M., Bailey, C.A., Rivers, S.E.: A framework for human-agent social systems: the role of non-technical factors in operation success. In: Savage-Knepshild, P., Chen, J. (eds.) *Advances in Human Factors in Robots and Unmanned Systems. AISC*, vol. 499, pp. 137–148. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-41959-6_12
19. Malle, B.F.: *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. MIT Press, Cambridge (2004)
20. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint* (2017). [arXiv:1706.07269](https://arxiv.org/abs/1706.07269)

21. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: beware of inmates running the asylum. In: IJCAI 2017 Workshop on Explainable AI (XAI), p. 36 (2017)
22. Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., Doshi-Velez, F.: How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. arXiv preprint (2018). [arXiv:1802.00682v1](https://arxiv.org/abs/1802.00682v1)
23. Neerincx, M.A., van Diggelen, J., van Breda, L.: Interaction design patterns for adaptive human-agent-robot teamwork in high-risk domains. In: Harris, D. (ed.) EPCE 2016. LNCS (LNAI), vol. 9736, pp. 211–220. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_22
24. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM, August 2016
25. Scheutz, M., DeLoach, S.A., Adams, J.A.: A framework for developing and using shared mental models in human-agent teams. *J. Cogn. Eng. Decis. Making* **11**, 203–224 (2017)
26. Staab, S., Studer, R. (eds.): Handbook on Ontologies. Springer, Heidelberg (2010). <https://doi.org/10.1007/978-3-540-24750-0>
27. Su, X., Matskin, M., Rao, J.: Implementing explanation ontology for agent system. In: 2003 Proceedings IEEE/WIC International Conference on Web Intelligence, WI 2003, pp. 330–336. IEEE (2003)
28. Taylor, G., Knudsen, K., Holt, L.S.: Explaining agent behavior. *Ann Arbor* **1001**, 48105 (2006)
29. Tiddi, I., d’Aquin, M., Motta, E.: An ontology design pattern to define explanations. In: Proceedings of the 8th International Conference on Knowledge Capture, 8 p. ACM (2015)
30. van der Waa, J., van Diggelen, J., Neerincx, M.A., Raaijmakers, S.: ICM: an intuitive, model independent and accurate certainty measure for machine learning. In: 10th International Conference on Agents and AI (ICAART 2018) (2018)
31. van der Waa, J., van Diggelen, J., Neerincx, M.A.: The design and validation of an intuitive certainty measure. In: Proceedings of IUI 2018 Workshop on Explainable Smart Systems (2018)
32. van der Waa, J., Robeer, M.J., van Diggelen, J., Brinkhuis, M.J.S., Neerincx, M.A.: Contrastive explanation for machine learning in adaptive learning (in preparation)
33. Wang, W.: Self-management support system for renal transplant patients: understanding adherence and acceptance. Ph.D. thesis. Delft University of Technology, The Netherlands (2017)
34. van Welie, M., van der Veer, G.C.: Pattern languages in interaction design: structure and organization. In: Proceedings of Interact 2003, 1–5 September, Zürich, Switzerland, pp. 527–534, IOS Press, Amsterdam (2003)



Crew Resource Management for Automated Teammates (CRM-A)

Robert J. Shively¹(✉), Joel Lachter¹, Robert Koteskey²,
and Summer L. Brandt²

¹ NASA Ames Research Center, Moffett Field, CA 94035, USA
{robert.j.shively, joel.lachter}@nasa.gov

² San Jose State University, Moffett Field, CA 94035, USA
{robert.w.koteskey, summer.l.brandt}@nasa.gov

Abstract. Crew Resource Management (CRM) is the application of human factors knowledge and skills to ensure that teams make effective use of all resources. This includes ensuring that pilots bring in opinions of other teammates and utilize their unique capabilities. CRM was originally developed 40 years ago in response to a number of airline accidents in which the crew was found to be at fault. The goal was to improve teamwork among airline cockpit crews. The notion of “team” was later expanded to include cabin crew and ground resources. CRM has also been adopted by other industries, most notably medicine. Automation research now finds itself faced with similar issues to those faced by aviation 40 years ago: how to create a more robust system by making full use of both the automation and its human operators. With advances in machine intelligence, processing speed and cheap and plentiful memory, automation has advanced to the point that it can and should be treated as a teammate to fully take advantage of its capabilities and contributions to the system. This area of research is known as Human-Autonomy Teaming (HAT). Research on HAT has identified reusable patterns that can be applied in a wide range of applications. These patterns include features such as bi-directional communication and working agreements. This paper will explore the synergies between CRM and HAT. We believe that HAT research has much to learn from CRM and that there are benefits to expanding CRM to cover automation.

Keywords: Crew Resource Management (CRM)
Human-Autonomy Teaming (HAT) · Automation

1 Why Human-Autonomy Teaming (HAT)?

For centuries, automation has been making humanity more productive. Historically this automation has chiefly replaced the brawn of human labor, leaving people as the brains. Recently, however, machines have been used by humans for increasingly cognitive tasks. Computers can play chess better than humans. They can search the web to

The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.

answer your questions. They can drive cars. But automation is not perfect. It is brittle, breaking, often catastrophically, when taken out of the “comfort zone” it was designed for. It has crashed stock exchanges [1] and cars [2]. Self-driving cars still need to “phone home” when encountering a person directing traffic [3]. Aviation has not been immune to this increase in automation, from relatively simple autopilots to sophisticated flight management systems. These innovations have not always been smooth [4]; however, the development of automation as tools to improve flight safety and efficiency has continued.

As the nature of automation changes, so should the role of the human when interacting with automation. Automation is moving from the realm of simple “tools” into intelligent cognitive agents that can function as teammates, similar to human teammates. Innovations in artificial intelligence as well as increases in the speed and memory of the underlying hardware have spurred this shift. Cognitive agents can now suggest courses of action, monitor the operator actions and physiology, and monitor the environment. With all of these new capabilities, work has begun to investigate how the relationship of the human and the automation can be shifted toward that of teammates to best take advantage of this phenomenon. This idea of Human-Autonomy Teaming (HAT) goes beyond simply giving a person a computer. The idea is that humans and automation should work together the way (well-functioning) human teams do, bouncing ideas off of one another, backing each other up when they sense potential problems, and keeping each other informed of what they are doing. HAT is currently recognized as a promising solution to the problems of human operators managing increasingly complex work systems. A human-autonomy team has been defined as “the dynamic, interdependent coupling between one or more human operators and one or more automated systems requiring collaboration and coordination to achieve successful task completion,” [5] a definition that has been picked up by others [6, 7]. As such, it is being developed and pursued in many operational areas such as robotics [8], commercial aviation [9], and UAS operations [10].

Aviation, with its very systematic approach to safety, may prove to be both a source of inspiration for developing better human-autonomy teaming, and an industry where its benefits might be most profitably exploited. While computers have eliminated the need for the navigator and (together with an increase in reliability) the flight engineer, regulatory barriers have slowed the ascent of automation in the cockpit. Years of cultivating teamwork on the flight deck have contributed to aviation’s superb safety record. Here we explore potential synergies between the lessons the airline industry has learned from years of studying teamwork and this newer field of human-autonomy teaming. Can we make human-autonomy teaming more effective by looking at the aviation model? Can we introduce advanced automation to the flight deck more safely if we introduce it as a team member? In the cockpit, where the importance of teaming has long been understood, the skills necessary for good teaming and the training procedures for teaching those skills have been codified under the umbrella of Crew Resource Management (CRM). Here we explore the synergies between HAT and CRM that may allow these new computerized teammates to be incorporated into CRM and may improve HAT concepts by incorporating the lessons learned from the development of CRM.

2 What Is Crew Resource Management (CRM)?

2.1 History of CRM in Human Teams

What can we learn from CRM as it has been developed for human teams? Although the discipline has expanded to other high-risk high-reliability industries, the multi-crew airline flight deck is where CRM has its roots.

Risk management has always been the core task for an airline pilot. In the early days of transport aviation, risks due to mechanical failure were more prominent and aircraft were less reliable. The majority of threats were clear and external to the human who was tasked with operating the machine. Early airline captains were solo performers whose technical skills were sharpened by absolute necessity. Their selection, their environment, and their culture reinforced strong, independent personalities and isolated individual decision making. As large aircraft became more complex, a single pilot could no longer operate the aircraft. The crew compliment grew but was comprised of people who were still focused on individual tasks. As the job became a more team-oriented endeavor, flight deck culture and nature of the individual pilot had not changed to reflect this shift.

By the late 1960's and 1970's transport accidents due to mechanical failure had drastically decreased. Advances in aviation technology like jet engines, modern avionics, and increased automation, so lowered the accident rate that the majority of new occurrences were now found to be crew related. Thus, further safety improvement could most easily be found from within the human team [11]. Landmark accidents like UAL 173, caused by fuel exhaustion, and EAL 401, brought down by the distraction from a faulty landing gear indicator, highlighted the need for training on team leadership and decision making for multi-crew aircraft [12]. The rugged, isolated individual was perhaps not the ideal model for what was now clearly a team activity.

A breakthrough study, conducted by Ruffel Smith [13] suggested a correlation between the leadership and communication style of the captain with overall crew performance. Further research reinforced this hypothesis [14, 15], and during the 1980's, industry and government came together to form what was at the time, a new discipline of aviation team training called Cockpit Resource Management, or CRM. As it evolved and became recognized as applicable to the larger aviation community, it was renamed Crew Resource Management. CRM has been defined as "using all available resources – information, equipment, and people – to achieve safe and efficient flight operations" [16, p. 20].

In the 1990's, CRM training programs were introduced at major airlines and in military aviation units. Many of the concepts originally trained were lifted from business school management training templates and were not particularly well received by pilots. The topics in these classroom seminars typically included assessments of personality and leadership style. While these assessments might have been good predictors of performance, they did not necessarily allow for actual behavioral change. The goal of the training was an attitudinal shift, but this was hard to measure and equally difficult to reliably move to the flight deck.

Over the course of two decades, emphasis has shifted to identifying observable flight deck team behaviors (good and bad) that could be trained and evaluated in actual

line operations. Task analyses were conducted at major carriers and research organizations. Both technical and CRM skills were identified and then used to create training in high fidelity, full flight, line-oriented simulations. These Line Oriented Flight Training (LOFT) events are now a primary means of introducing, reinforcing, and evaluating CRM skills [17]. CRM training today still includes an indoctrination seminar for new hires, but now CRM skills are embedded in all aspects of multi-crew transport training. They are trained and evaluated alongside technical skills at every training opportunity. Current CRM doctrine uses the concept of Threat and Error Management (TEM). This paradigm seeks to acknowledge the universal existence of human error and outside threats. TEM seeks to engage the team in actively searching for those threats and inevitable errors, then through use of CRM skills, to avoid, minimize or correct them.

The goal of CRM is the optimization of the human team. Whether the task is performing a checklist, monitoring a standard operating procedure, leading the crew, or mindfully following the direction of the leader, the whole point is for the members of the team to form, and perform optimally, and in synergy. We will now look at recognized CRM concepts and tools that are trained to, and used by, human teams in aviation, with the intent of adapting concepts from this field to the optimization of the new human-automation team.

2.2 Generally Recognized CRM Training Concepts for Human Teams

Threat and Error Management. As noted above, Threat and Error Management is one recognized framework used to convey a common mental strategy for identifying potential threats to safe operation, for identifying potential errors, and for correcting them when they inevitably happen [17]. TEM defines a continuum of safety ranging from safe operations on one side to an undesired equipment state on the other (i.e., a series of errors has occurred and gone unrecognized and unmitigated so that an unsafe condition or accident has occurred). The goal of TEM is to always remain in, or return to, safe operations while avoiding an undesired equipment state.

Avoid. Teams are taught that all operations have certain inherent threats. A threat can be internal or external to the team. These are identified and called out during briefing with the intent that this affords a greater likelihood that these dangers will be consciously mitigated by all members of the team. These known threats are to be avoided and a strategy for doing so can be articulated even before the team begins its task. Common threats might include a new or inexperienced team member (internal), or a hazardous weather condition (external).

Repair. An important new implication of this model is that the team expects that it will make errors during the course of its action. In aviation training this was an innovation. Previous generations of aviation professionals were expected to perform procedures and tasks to near perfection. Individuals were likely to minimize or discount errors both in training and actual practice because there was professional and social pressure to do so. In the TEM paradigm, individual errors are expected, so team members may be more likely to call out their own errors and those of others with less stigma attached.

While individuals are expected to make errors, the team as a whole is expected to respond to those and correct them before they lead to an undesired equipment state.

Recover. Should the first two levels of defense fail to capture the inevitable errors and threats that a team is likely to encounter, an undesired equipment state may occur. The task then becomes recovering the operation to a safe state. It is expected that CRM skills be utilized to return to a safe operation.

Verbalize Verify Monitor. One CRM tool used to capture and mitigate error is known as Verbalize Verify Monitor or VVM. Using VVM, the team member planning on taking an action first verbally states the intention of the action. This allows both the acting member and a monitoring team member to verify that the action matches the intent of current team goals. If that test is met, the action is carried out and the results can then be effectively monitored.

Verbalize. A statement that focuses the team on the accuracy of the next action to be taken. This need not be a simple statement. Often, a question is appropriate. Especially if the initiator is checking the accuracy of their own perception. As an example: “I intend to turn right.” or “Is the next turn to the right?” may both be appropriate, depending on the verification goal of the person taking action.

Verify. When an intent is stated, another team member is needed to verify the accuracy of the statement. The verification will usually also be stated but may be passed by non-verbal means if appropriate to the workload and situation. The important aspect is that another member is now engaged and their input is received and acknowledged.

Monitor. This last step requires continued engagement of the team to ensure that the intent just stated and verified is actually carried out accurately. In the case of automation management, this is particularly useful during multi-step changes or complicated procedures. A simple aviation example might involve the pilot flying calling out an intended automation mode change, executing the change after ensuring the monitoring pilot is engaged, and then both pilots monitoring the results of the change for accurate results. This process is used not only in automation management, but in all aspects of team action. For instance, a flight deck team taxiing on an unfamiliar airport would use VVM to ensure the aircraft does not deviate from its cleared route (i.e. enter an undesired equipment state). Using VVM, the pilot steering would proactively call out each turn and holding point well before the action must be taken. This communicates intent, focuses the monitoring team member, and allows for checking of common situation awareness. VVM works to verify technical as well as non-technical tasks. It may even be used during briefing to verify and align team goals, or as a tool to engage low-response team members.

Standard Operating Procedures. A common observation made concerning airline crews is that it seems remarkable that persons who have never met each other could come together in a very short time, form a team, and then operate a complex piece of equipment in a challenging environment. This seemingly remarkable ability lies in the training of those people to the same set of standard procedures. They have a reasonable expectation of what skill level they will encounter, and indeed, after the advent of CRM, even an expectation of how they will be treated as the team forms and acts.

Well known human factors tools like checklists, standardized callouts, and stabilized approach criteria all fall into the category of Standard Operating Procedures (SOPs). Some of these tools have been around as long as aircraft have existed and are used in many other high-risk high-reliability activities. SOPs ensure that a known, safe, efficient set of actions is used to navigate through complex procedures that require great accuracy. At an airline, they are developed by expert teams and vetted through internal and regulatory processes to verify their worth and efficacy.

SOPs are intended to be the familiar landmarks that provide reinforcement, guidance, and reassurance during the progression of the project that is the flight of a transport aircraft. Until the advent of CRM, there was no SOP for the operation of the human team. Even with an extensive set of well-defined procedures for operation of the equipment, it is possible for a group of humans to fail in the execution of those procedures if they cannot work together. CRM, then, can be thought of as set of SOPs for operating in a team with other humans. When used by high-functioning teams employing a full pallet of CRM skills, technical SOPs are landmarks that appear as familiar signposts that are the culmination and verification of various team actions. Conversely, to low-functioning teams without a good grasp of solid CRM skills, these landmarks may come as a surprise, presenting last-chance safety backstops, rather than the mile markers of a safe operation.

Systems Approach to Training CRM Skills (AQP and ISD). It is important to consider how the different needs and existing skills of a particular organization influence the CRM training product. How does an organization identify the particular CRM skills and values it would like to emphasize, and how does the organization evaluate whether or not the program is effective? For airline operations, this process is codified by the FAA in two Advisory Circulars. One, FAA Advisory Circular 120-51E, concerns CRM [18], the other FAA Advisory Circular 120-54A, concerns the Advanced Qualification Program or (AQP) [19]. The CRM publication outlines basic CRM training topics, and the AQP publication describes a systematic process for training and evaluating CRM skills alongside technical skills.

For airline operations using an AQP, the requirements attempt to ensure that CRM skills and the resulting training program are tailored to that airline's needs. A systematic approach is advocated called Instructional Systems Design (ISD). This model starts with needs assessment to see where the organizational baseline is. A training goal is then defined and objectives written to support that goal. Once training is written and delivered, data is gathered to facilitate a process of continual improvement.

2.3 Common Target Concepts for CRM Training

In the airline industry, major airlines generally still present a CRM indoctrination seminar that includes some aviation human factors and an introduction to CRM general concepts. This establishes that common vocabulary which is used for training and evaluation of the concepts and tools throughout a pilot's career at a given carrier. This involves learning to operate the "human equipment" one is teamed with at the same level of proficiency as the mechanical equipment in use. Maximizing the team's efficiency and output is seen as being of equal importance to technical proficiency.

How is this done quantitatively and what do those behaviors look like? An excellent summary of recognized CRM skills that might be generally trained to new hires and evaluated in experienced operators is found in the book *Crew Resource Management* [20]. Below, we briefly lay out the CRM skills they identified. We will then discuss the synergy these CRM skills have with HAT and the potential benefits of expanding CRM to cover automation.

Communication. Making sure that there is bottom up communication as well as top down is the core of all other CRM competencies. CRM training teaches techniques for clear communication. One example is the use of active listening, which is the mindful repetition of the sender's message back to them as a check for understanding. Another is teaching common, simple communication models to team members. An example of a model that might be trained during a CRM seminar is a three step process in which communication has not occurred until 1. A message is transmitted, 2. The message is received, and 3. Feedback is provided. Proactive communication utilized mindfully by all team member is essential for all other aspects of CRM to work.

In a high functioning human team, a person who has important information makes sure that information is communicated. This is true whether that team member is a decision maker or a subordinate. Alternate modes and channels are attempted until it is clear that the communication has occurred. As human teams form, they become more efficient in their communication. They begin to learn how to communicate with the individuals they are teamed with and then adapt their style to fit the person and the situation.

Briefing. Dedicated briefings are useful in organizing teams and maintaining a common plan so that actions are properly choreographed. A team leader may use the preflight briefing to set the tone of team interaction in addition to the simple passing of pertinent information. During flight, a pause to re-brief as conditions and goals change is also useful in quickly redirecting or re-focusing the team. Post flight debriefing is important in the continuing process of improvement for individuals, for teams that are likely to reform, and for the organization as a whole.

Backup Behavior. While briefings and SOPs set out roles and responsibilities for each crew member, in well-functioning teams, task allocation is not absolute. Particular circumstances may result in one crew member's assigned duties exceeding their workload capacity. Under such circumstances, workload should be shifted so that it is balanced across team members. This kind of transition can be seen during an off-nominal flight deck event requiring a shift from nominal SOP duties to alternate duties that compensates for the increased task complexity and better distributes the changed workload. For instance, the captain may assign the first officer both pilot flying duty and the task of communicating with ATC, a shifting of roles as the flying pilot would normally rely on the pilot monitoring for this task. Because of the off-nominal (perhaps requiring someone to run a checklist and work on a change of destination), workload must be re-distributed.

Mutual Performance. "To err is human." CRM attempts to prevent human errors from resulting in incorrect actions through mutual monitoring among crew members. Crew members are taught to give and take advice in an open and non-judgmental

manner. An important aspect of this is the VVM technique discussed in Sect. 2.2 above. Monitoring is also an important aspect of Backup Behavior. Crew members are taught not only to watch for individual lapses in judgment, but also their partner's overall workload and mental state, and to offer greater assistance when a partner becomes overloaded.

Team Leadership. While many of the CRM skills have the effect of making teams more egalitarian, there is still a recognized need for one person to be in charge. The desired goal of the leader is to ensure that all team members are used optimally, and that they are engaged both with the task and with the team. Good leaders organize the team in a way that makes appropriate use of each team member's skill and ability and keeps them working together in a positive manner. For example, a good captain might have someone with more flight experience as a first officer, or might have a rookie. What is appropriate monitoring and mentoring in one case might be perceived as condescending and micromanaging in the other. It is the leader's task to make this type of assessment and modify their interaction to obtain maximum results from the team.

Decision Making. A key goal of CRM is to improve decision making by encouraging the consideration of multiple possible courses of action and assessing each using as complete a collection of information as possible. To do this, crew members are taught to bring up potentially relevant information and alternative actions, considering the possible consequence of each, with an attitude of "what is right, not who is right."

Task-Related Assertiveness. In order to maintain a collaborative decision making atmosphere it is important that team members develop an appropriate level of assertiveness. They must be able to communicate information and suggestions with persuasive logic while maintaining an ability to listen and be persuaded by other team members. This helps ensure that all available information is put on the table in a transparent manner when decisions are being made. In fact, it goes beyond this, to require team members to bring information forward even when it is not explicitly requested. Appropriately assertive input, particularly from subordinates, is key to sound communication. It is a cornerstone of good followership, the important but often neglected obverse to good leadership.

Team Adaptability. Humans show a strong natural tendency to maintain a particular course of action, even when, from a purely logical standpoint, it no longer makes sense to do so (e.g, the sunk cost fallacy [21]). CRM attempts to counter this tendency, by encouraging continuous re-evaluation of the current course of action, recognition of possible threats to the current goals, and discussion of options.

Shared Situation Awareness. Many of the skills discussed in this section can be viewed through the lense of developing an accurate shared awareness of the situation. Communication and briefings serve to keep crew members on the same page, with shared goals and a shared understanding of the environment so that they do not act at cross purposes. CRM skills related to monitoring, assertiveness and adaptability serve to maintain the accuracy of this shared understanding.

3 CRM for Human-Autonomy Teaming

Automation research now finds itself faced with similar issues to those faced by aviation 40 years ago: how to create a more robust system by making full use of both the automation and its human operators. This section will examine two overarching HAT concepts that have been proposed and then looks at how the CRM skills identified above might be applied to a human-automation team.

3.1 HAT Concepts Supporting CRM-Like Behavior

Bi-directional Communication. As with the VVM pattern in human teams, it is important to develop a style of communication that makes sure that information is communicated across all team members in human-autonomy teams. This pattern has been referred to as bi-directional communication [22]. For automation to participate as a teammate, it is critical to have a bi-directional communications channel. This will allow humans to team effectively with automation and allows the human (and automation) to question, share hypotheses, provide additional input, etc. just as human teammates would. This bi-directional communication is critical and enables many of the CRM elements to follow. It must be bi-directional to allow the pilot to input information into the system that the automation might not have access to via sensors or databases. For example, when deciding on an alternate airport due to a medical emergency, the pilot might know more about the medical facilities in and around particular airports than does the automation. For the automation to participate fully as a partner, it needs to share this information, therefore this channel needs to exist. Similarly, the automation needs to be able to alert, share hypotheses, level of confidence, etc. with the human teammate. This then, allows the human to better judge the value and understanding of the automation and trust appropriately. This is an example of transparency, being used to calibrate trust, enabled by bi-directional communications. Teammates often discuss options, brainstorm on solutions and openly discuss courses of action. For automation to be on a team, this bi-directional communication needs to exist. Bi-directional communication is key to solving a number of the issues typically found in highly automated systems. Bi-directional communication can make systems more transparent and less brittle and further can facilitate intent based interface design.

Working Agreements. In human-human teams, SOPs provide both a level of predictability for how team members will react in a variety of situations, along with the ability to plan for many situations offline. Work in HAT has developed a concept similar to SOPs for use with automation [23, 24]: working agreements. Working agreements encapsulate goals, procedures, and division of responsibility into a package that can be specified offline and instantiated quickly in real-time situations. Working agreements specify who (automation or human, and, in the case of humans, which human) is responsible for performing various acts associated with a particular situation. This responsibility can be conditional. For example, the automation might be given autonomous authority to follow a route unless potential hazards are detected, at which point it might alert the human operator for verification that it should proceed.

3.2 Developing CRM Skills with Autonomy

As discussed in Sect. 2.2 above, human teams are taught certain CRM skills that provide standardized mechanisms for using the situation specific SOPs to improve performance on certain measures thought to improve operational outcomes. Here we discuss how automation designed to work with bi-directional communication and working agreements can mirror those same skills.

Communication. CRM in the context of human teams emphasizes the need for communication to flow both up and down the chain of command. Our bi-directional communication pattern is designed to enable something similar between humans and automation. Part of the reason for CRM was that superiors do not always want to hear what their subordinates are saying and subordinates are often scared to speak-up. Analogous problems may occur between human and automation. While we can assume automation will not fear speaking up, how can we make sure that the human listens appropriately? Some research indicates that manner in which information is communicated influences the degree to which operators accept and rely on automation [e.g., 25]; however more work is needed in this area.

Further, just as humans must adapt their communication styles to their teammates, it may be appropriate to build similar adaptability into the automated team members. Just as humans must be trained to operate synergistically within their teams, perhaps automation should also be capable enough to recognize individual human style in order to maximize HAT performance. How will the automation provide a metaphorical “touch on the shoulder” when its human partner is deemed not to be listening?

Briefing. As with human-human teams, human-autonomy teams must share a common plan to assure that actions are properly choreographed. This goes beyond the flight plan to include alternate airports (depending on flight progress), weather, aircraft status, any potential areas of concern, and any issues that would normally be discussed with the crew. Digital representation of the flight plan exists and so is straightforward to transfer to the automation; however, other aspects may be harder to transmit. It may be necessary to build a “briefing interface” on top of the bidirectional communication channel to allow the crew to easily and fully provide this information to the automation. Such an interface would be a logical place to define working agreements between the crew and the automation.

Here may be an ideal opportunity to build the human-automation team in much the same way as human-human teams are built. A briefing for a human team provides not only information transfer, but, importantly, affords the team an opportunity for each member to adjust team dynamics and to begin forming communication strategies.

It is also important to consider that briefings do not only occur at the beginning of a task. High functioning crews use them to re-focus the team when the situation has changed and at the end of the mission to review lessons learned, both good and bad. These other modes of briefing may have utility in helping humans and automated team members align goals and adapt to each other.

In addition, humans also use briefings to assess one another and to modify their behavior to better conform to each teammate’s preferred style of communication or operation. Could the briefing opportunity for the automation to gather information on

the human team member's preferences? Perhaps the human team member might proactively push this preference information to the automation as part of the briefing. One could perhaps envision an opportunity for carrying this information in a profile of some kind that could follow a human team member from station to station.

Backup Behavior. SOPs help human team members anticipate the needs of others by providing clarity about each other's responsibilities. CRM training teaches pilots formal ways to modify these responsibilities by shifting workload between members to create balance during periods of high workload or pressure while maintaining a clear understanding of who is responsible for what. Similar flexibility can be built into working agreements between humans and automation. As noted above, a key feature of working agreements is the ability to specify the conditions under which each party is responsible for taking certain actions. Stress and workload levels can be among these conditions. However, doing so requires that these levels can be sensed or conveyed to the automation without adding to the overall workload. Sensing raises its own issues; how does the operator know that the automation has sensed high workload and changed the task allocation? An alert or annunciation system would have to be very sophisticated to avoid distracting the operator in a high workload situation. Operators could initiate such changes vocally or with a simple interface such as a button or dial. This would allow the operator to control the timing of any task reallocation and assure that his or her situation awareness is maintained. Again, this is an opportunity for further research.

Mutual Performance Monitoring. An important aspect of CRM is crew members monitoring the performance of other crew members. This is also true for automation. The automation needs to be able to monitor the crew: Are tasks being performed in a timely manner? Is the pilot planning and staying "ahead of the aircraft"?

As in the CRM construct of TEM, errors should be expected, recognized, and mitigated as necessary. Humans find a variety of ways to begin monitoring each other's actions. Indeed, they start to monitor the quality of team interaction as well. In high-functioning human teams, members quickly learn each other's style. Even on a short mission, with formerly unfamiliar teammates, this information about how the other operates may provide enhanced monitoring ability. HAT monitoring of error and compliance may benefit from perception of, and adjustment to, these individual human characteristics and preferences.

An extreme case of required monitoring would be for nefarious behaviors. In human teams, much of this kind of screening is accomplished as the team forms and briefs. Under HAT, we would expect that under most circumstances, the human would lead the team. However, if the pilot has significantly deviated from the flight plan without a plausible explanation, the automation may have the authority to take certain actions: contact air traffic control, company dispatchers or in extreme cases, take control of the aircraft. In addition, it may be necessary for the automation to monitor pilot physiological state: heart rate, blood pressure, eye gaze, etc. to verify that the pilot is fit and operating in an acceptable physiological state. This could become critical in the case of incapacitated pilots; the automation may have the authority to take control in these situations. For either of these types of monitoring to occur, the automation needs insight into the actions on the cockpit and physiological sensors on the pilot/crew.

Monitoring needs to be bi-directional in nature. That is, the pilot/crew must be able to monitor the automation just as the automation monitors the pilot. To do so, the pilot needs insight into the automation. Transparency into the processing and decision making provided by bi-directional communication is critical to this monitoring. When the automation alerts the pilot or offers suggestions (e.g., alternate airports), the logic of the processing needs to be available for examination by the pilot. But, the need goes even further; the pilot needs to have indications that the automation is monitoring and performing as intended. The bi-directional communication interface should be designed to provide this information.

Team Leadership. For the foreseeable future, the pilot (human) will be the pilot in command and therefore the leader of this team, and thus ultimately responsible for the performance of the team. However, automation can still fully participate as a teammate, as long as several of the attributes discussed thus far are in place: e.g., communications, monitoring, transparency. These will provide the leader with the mechanism and the information required to direct and coordinate the activities of team members, encourage team members to work together, assess performance, assign tasks, develop team knowledge, skills, and abilities, motivate, plan and organize, and establish a positive team atmosphere.

Coordination with the automation may be through working agreements. As with human teams, it is incumbent upon the leader to know the abilities of individual team members and understand how to communicate effectively with them. In HAT, this may look like extensive training for the human team leader on the automated team member's capabilities and limitations. In addition, it may be desirable for the automated team member to practice the analog of good followership as described for human teams. This might involve the automation retaining some capability to recall communication or interface preferences of the human team leader.

Decision Making. Good decision making in human teams involves gathering and integrating information, identifying alternatives, and considering the consequences of each alternative. CRM encourages options developed by one team member to be evaluated and refined by other team members. This suggests that for automation to have good CRM skills it should be able to do three things that current automation typically cannot: evaluate options proffered by a human operator, give reasoning behind options it proffers, and compare options.

Evaluation of options proffered by the operator might be facilitated with a "course of action scratch pad," that would allow the operator to input a proposed course of action (e.g., commands, reconfiguration, or routing) and have the automation evaluate it, presenting predicted outcomes (e.g., risk assessment, estimated fuel usage, ETA, etc). Similarly, when proffering a course of action, automation should be able to give similar evaluations along with an indication of what options were considered in developing this course of action.

Task-Related Assertiveness. A related measure of good CRM in human teams is the ability of members to communicate their ideas, opinions, and persuasively while remaining open to being convinced by the facts that other options are better. Working

agreements can be used to implement such task related assertiveness in an automated system. A working agreement can specify conditions under which automation should “speak up” by alerting human operators to problems with the current course of action and/or offering alternatives. In addition, several channels for gaining the human’s attention may be desired. In human teams, a physical touch, a change in verbal tone or cadence, or even specific standardized phrases are all methods currently in use to gain the attention of a crew member who is not attending to a particular message. Some of these may translate well to automation while others clearly will not.

Team Adaptability. In human teams, good CRM requires that the team be able to alter a course of action or adjust strategies when new information becomes available, rather than push forward with a suboptimal or even infeasible plan. Properly designed automation can help with this. While people are often reluctant to give up on a course of action once it is embarked upon, automation has no such limits as long as it is running open loop. Automation can detect when the current course has become sub-optimal and propose deviations. To prevent the automation from overwhelming the human operator with new options (e.g., modifying an aircraft’s trajectory every time the wind shifts slightly), working agreements can be developed that limit proposals such proposals to cases where the risk or cost difference meets a certain threshold. Alternatively, a working agreement could be developed that gives the automation authority to make small deviations from the current course of action autonomously.

Shared Situation Awareness. In human teams, it is important to maintain a common understanding of the task and team environment to keep everyone working toward a common goal. When new information becomes available it must be communicated or team members may find themselves working at cross purposes (e.g., if one pilot hears a controller say descend to FL270 and the other hears FL260). This issue is even more important in dealing with human-automation teams because the human and the automation do not innately have the same information available to them. The automation takes in data from various sensor feeds at a level of detail that, even if it is available to the human, the human cannot process. Similarly, the automation can make calculations much faster and more precisely than the human, allowing it to quickly recognize and react to changes in the environment. The automation may not, however have the full range of senses that the human has. This is particularly important when it comes to understanding other people who, in most cases, set the objectives of the system. The bi-directional communication channel discussed in Sect. 3.1 above will be very important in allowing both the human and the automation to integrate this information and assuring that these representations are compatible. In some cases, it may be appropriate to “re-brief” as mentioned above in order to check for common goals and understanding of the mission. One could envision that this action could be called by either the automated component (if it senses that the human is taking inconsistent action) or by the human member of the team.

4 Conclusion

CRM has become deeply integrated into airline crew training. As automation rises the level of a teammate, it is imperative that this new status be reflected in CRM curriculum. It is recommended that airlines review their CRM training and incorporate this new more powerful automation paradigm as a critical component. It is perhaps equally imperative that research into CRM be incorporated into the design of these new non-human team members. The human-automation team should be developed in such a way that the human team member may eventually trust and interact with the automated team member in many of the same ways as they would with another human.

References

1. Steiner, C.: *Automate This: How Algorithms Took Over Our Markets, Our Jobs, and the World*. Penguin Group, New York (2012)
2. Stewart, J.: Why Tesla's autopilot can't see a stopped truck. *Wired* (2018). <https://www.wired.com/story/tesla-autopilot-why-crash-radar/>
3. Lawler, R.: Nissan's SAM uses humans as a backup for self-driving tech. *Engadget* (2017). <https://www.engadget.com/2017/01/05/nissans-sam-uses-humans-as-a-backup-for-self-driving-tech/>
4. Wiener, E.L.: Cockpit automation. In: Wiener, E.L., Nagel, D.C. (eds.) *Human Factors in Aviation*, pp. 433–461. Academic Press Inc., San Diego (1988)
5. Cuevas, H.M., Fiore, S.M., Caldwell, B.S., Strater, L.: Augmenting team cognition in human-automation teams performing in complex operational environments. *Aviat. Space Environ. Med.* **78**, B63–B70 (2007)
6. Langan-Fox, J., Canty, J.M., Sankey, M.J.: Human-automation teams and adaptable control for future air traffic management. *Int. J. Ind. Ergon.* **39**, 894–903 (2009)
7. Strybel, T.Z., et al.: Measuring the effectiveness of human autonomy teaming. In: Baldwin, C. (ed.) *AHFE 2017. AISC*, vol. 586, pp. 23–33. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60642-2_3
8. Chen, J.Y.C., Barnes, M.J.: Human-agent teaming for multi-robot control: a literature review (ARL-TR-6328). Human Research and Engineering Directorate, Aberdeen Proving Grounds, MD (2013)
9. Brandt, S.L., Lachter, J., Russell, R., Shively, R.J.: A human-autonomy teaming approach for a flight-following task. In: Baldwin, C. (ed.) *AHFE 2017. AISC*, vol. 586, pp. 12–22. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60642-2_2
10. Fern, L., Shively, R.J.: A comparison of varying levels of automation on the supervisory control of multiple UASs. In: *Proceedings of AUVSI's Unmanned Systems North America*, Washington, DC (2009)
11. Kanki, B.G., Helmreich, R.L., Anca, J.: *Crew Resource Management*, 2nd edn. Academic Press, San Diego (2010)
12. Helmreich, R.L., Foushee, H.C.: Why CRM? Empirical and theoretical bases of human factors training. In: Kanki, B.G., Helmreich, R.L., Anca, J. (eds.) *Crew Resource Management*, 2nd edn, pp. 3–57. Academic Press, San Diego (2010)
13. Ruffel Smith, H.P.: A simulator study of the interaction of pilot workload with errors, vigilance, and decisions, NASA-TM-78482. NASA Ames Research Center, Moffett Field (1979)

14. Chidester, T.R., Helmreich, R., Gregorich, S., Geis, C.: Pilot personality and crew coordination: implications for training and selection. *Int. J. Aviat. Psychol.* **1**, 23–42 (1991)
15. Chidester, T.R., Kanki, B.G., Foushee, H.C., Dickinson, C.L., Bowles, S.V.: Personality factors in flight operations. Volume 1: leader characteristics and crew performance in a full-mission air transport simulation. NASA Ames Research Center, Moffett Field, California (1990)
16. Lauber, J.: Resource management in the cockpit. *Air Line Pilot* **53**, 20–23 (1984)
17. Federal Aviation Administration: Flightcrew member line operational simulations: line-oriented flight training, special purpose operational training, line operational evaluation. Advisory Circular 120-35D (2015)
18. Federal Aviation Administration: Crew resource management training. Advisory Circular 120-51E (2004)
19. Federal Aviation Administration: Advanced qualification program. Advisory Circular 120-54A (2006)
20. Shuffler, M.L., Salas, E., Luiz, X.F.: The design, delivery and evaluation of crew resource management training. In: Kanki, B.G., Helmreich, R.L., Anca, J. (eds.) *Crew Resource Management*, 2nd edn, pp. 205–232. Academic Press, San Diego (2010)
21. Kahneman, D.: *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York (2011)
22. Shively, R.J., Lachter, J., Brandt, S.L., Strybel, T.Z.: Human-autonomy teaming in a flight following task. In: NATO HFM 247 Technical Activity Description (in press)
23. Miller, C.A., Parasuraman, R.: Designing for flexible interaction between humans and automation: delegation interfaces for supervisory control. *Hum. Factors* **49**, 57–75 (2007)
24. Gutzwiller, R.S., Espinosa, S.H., Kenny, C., Lange, D.S.: A design pattern for working agreements in human-autonomy teaming. In: Cassenti, D.N. (ed.) *AHFE 2017*. AISC, vol. 591, pp. 12–24. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60591-3_2
25. de Visser, E.J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., Parasuraman, R.: The world is not enough: trust in cognitive agents. In: *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting*, pp. 263–267 (2012)



An Integrated After Action Review (IAAR) Approach: Conducting AARs for Scenario-Based Training Across Multiple and Distinct Skill Areas

Lisa Townsend¹(✉), Joan Johnston², William A. Ross³,
Laura Milham¹, Dawn Riddle¹, and Henry Phillips¹

¹ Naval Air Warfare Center Training Systems Division (NAWCTSD),
Orlando, FL, USA

{lisa.townsend, laura.milham, dawn.riddle,
henry.phillips}@navy.mil

² Army Research Laboratory Human Research and Engineering Directorate
(ARL-HRED), Orlando, FL, USA

joan.h.johnston.civ@mail.mil

³ Cognitive Performance Group, Orlando, FL, USA
bill@cognitiveperformancegroup.com

Abstract. Due to resource constraints, labor intensive scenario-based training solutions often include training on more than one skill area consisting of distinct multiple learning objectives. However, After Action Reviews (AARs) taking place after training have not adapted and have either become complex and drawn out to accommodate more skill areas or worse, critical objectives are simply left out because there is no time left to cover them. These AAR challenges should be addressed because each skill area and objective should be discussed for optimal learning and team performance improvements to occur. An Integrated AAR (IAAR) approach designed to cover multiple skill area objectives can enhance scenario based training opportunities without encumbering a team member's ability to learn. During the Squad Overmatch (SOvM) training effectiveness evaluation different resources were developed to conduct an IAAR crossing multiple skill areas. Some of the resources developed worked well while others required revisions. The SOvM IAAR process and approach is described, lessons learned are discussed, and a new concept for an IAAR dashboard is presented.

Keywords: Scenario-based training · After action reviews (AARs)
Team training

The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.
Disclaimer: The views expressed herein are those of the authors and do not necessarily reflect the official position of the organizations with which they are affiliated.

1 Introduction

An AAR is a structured review or debrief process for analyzing differences between actual and expected performance after military training exercises or actual tactical events. AARs also provide a process for identifying and using lessons learned to improve tactical performance or change individual behaviors following scenario-based training. According to Army Doctrine Reference Publication (ADRP) 7-0 an AAR is "...a guided analysis of an organization's performance, conducted at appropriate times during and at the conclusion of a training event or operation with the objective of improving future performance. It includes a facilitator, event participants, and other observers. Team members, or participants, provide responses to questions about what happened, why it happened, and agree on how to sustain strengths and improve performance. Often, a team leader directs an AAR and focuses on only what could have been done better, paying little attention to what was done well and why. Formal AARs were originally developed by the U.S. Army in response to the need for arriving at performance improvements by blending squad member inputs with objective performance measures. Effective AARs are usually centered on formative feedback, self-monitoring and self-reflection, which can deepen and expand learning [1, 2]. An AAR is essentially an opportunity to improve tactical performance.

The Integrated After Action Review (IAAR) was developed by the Squad Overmatch (SOvM) research program- a multi-year, joint US Army – US Navy research effort - to improve individual and team performance under stressful conditions. SOvM training integrates tactical skills and team behaviors in five skill areas to improve mission effectiveness [3, 4]. The integrated training approach includes classroom training (knowledge acquisition), participation in simulation-based training (opportunity to practice what was learned in the classroom), and participation in live training (opportunity to apply what was practiced virtually and learned in the classroom). SOvM scenario based training includes an IAAR after each virtual and live scenario. The IAAR is introduced as an AAR that covers multiple skill areas that are integrated through scenario based training and discussed during an IAAR. Because training resources are limited, it makes sense that training objectives should be collectively combined for training (when resources can be shared and it is complementary to the skills areas to do so). An important difference between the IAAR and traditional AARs is where the discussions are focused. During a SOvM IAAR, the squad shifts its focus to teamwork behaviors, instead of predominantly focusing on tactical skills. The IAAR creates an atmosphere where each squad member's role shifts from Soldier being corrected to Soldier offering self-correction.

Both AARs and IAARs are opportunities to improve performance through facilitated discussions that start with agreement on an overall goal and training requirements. Each compares expected performance to actual performance and requires individual accountability for task performance. The main difference between an AAR and an IAAR is that an effective IAAR emphasizes collective learning across multiple skill areas (vs only tactical skills) and requires all squad members' (from the lowest level up) participation and engagement (vs the team leader doing most of the talking). IAARs that address tactics and teamwork require members to be accountable for team

performance and contribute to solutions and goals. Therefore, in order for an IAAR to be effective it should create a learning environment that provides opportunities for knowledge exchange, facilitates changes in behaviors, and is resourced to learn from information collected during scenario based training.

2 SOvM Training Effectiveness Evaluation

A Training Effectiveness Evaluation (TEE) of SOvM was conducted in June 2016 at Fort Benning, GA. It was led by the Program Executive Office for Simulation, Training, and Instrumentation, Army Research Laboratory Human Research and Engineering Directorate, Naval Air Warfare Center Training Systems Division, The MITRE Corporation, and Cognitive Performance Group. The U.S. Army Maneuver Center of Excellence, Maneuver Battle Lab, Clarke Simulation Center, and the McKenna training complex. These organizations provided the training and simulation resources at Fort Benning, GA.

Participants included four squads from the 82nd Airborne Division (Fort Bragg, NC) and four squads from the 75th Ranger Regiment (Fort Benning, GA). Each squad was augmented with a 68 W medic from the 690th Ground Ambulance, 14th Combat Support Hospital (Fort Benning, GA). Squads size ranged from eight to ten members. Four squads participated in an experimental condition and four squads participated in a control condition.

Squads in the experimental condition received classroom training, participated in two simulation-based training scenarios, participated in three live training scenarios, and engaged in an IAAR after each scenario. Control condition squads participated in only two live training scenarios and participated in a traditional AAR after each scenario.

Squads in the experimental condition received instruction from five instructors in five skill areas:

- Tactical Combat Casualty Care (TC3) – Trains communication and team member roles and priorities in response to medical tactical situations.
- Advanced Situation Awareness (ASA) – Trains human behavior pattern/threat recognition and decision making in complex environments.
- Resilience and Performance Enhancement (RPE) – Develops squad member skills in maintaining tactical effectiveness under combat stress.
- Team Development (TD) – Develops teamwork skills including Information Exchange, Communication Delivery, Supporting Behavior, and Team Initiative/Leadership.
- IAAR – Develops an understanding of the IAAR process, skills in applying the Force of Four framework, and methods for identifying the characteristics of an effective IAAR.

The cornerstone of SOvM training is the IAAR. It is the culminating event that provides the foundation for the integration of the skill areas and offers the opportunity for teams to detect errors, reflect on behaviors, and self-correct their performance. These activities lead to improved team performance.

3 Preparing for and Conducting the SOvM IAAR

IAAR preparation includes a variety of techniques to observe and collect examples of skill area behaviors and to discuss them openly. This approach is based on years of team research findings. The research found that participant feedback is most effective in improving performance [1]. These improvements occur when the team recognizes its less than optimal behavior, acknowledges the consequences of that behavior, generates solutions, and sets goals to improve behavior. Each person on the squad actively participates in the process by identifying examples of good and poor performance during scenarios and by contributing to opportunities where the team recognizes team errors and discusses more effective solutions. This approach encourages the team to collaborate on improving its performance through goal setting.

The SOvM IAARs for the TEE included the skill area instructors, an IAAR Facilitator (for SOvM, the Facilitator was the squad's Platoon Leader), and the Army squad itself. The IAAR followed a process with specific steps (see Fig. 1. 'IAAR Process' below).

The IAAR model includes an instructor for each skill area and an IAAR Facilitator who guides the discussion. These individuals observe and collect squad performance data during virtual and live scenario-based training. Then, during the IAAR that follows each scenario the Facilitator reviews performance objectives and elicits squad inputs about the tactical timeline. Then instructors review skill area learning objectives and ask squad members (1) where they struggled and excelled (triggers); (2) to agree on what went wrong and right (teamwork behaviors); (3) to propose a workable solution (identify correct procedure); and (4) to discuss real world outcomes and consequences. This is called the Force of Four, which provides a framework for team self-correction during the IAAR. With the support of the Facilitator and instructors, squads also set goals and integrate them into the next mission's planning. In this IAAR process, the Facilitator and instructors act as guides to keep the IAAR on track. Squad members contribute and engage in team self-correction across the integrated skill areas.

For this multiple skill area focused IAAR to work optimally a number of resources were developed and used to keep the IAAR on track and covering all the required objectives within each skill area within a 30–40 min timeframe.

For the SOvM TEE, job aids for Skill Area Observation and Assessment were provided so that each instructor could link skill area objectives to specific scenario events and injects, making it easier to identify whether specific behaviors occurred. These job aids were paired with individual Skill Area Scenario Event Timelines and Overlays (see Fig. 2 'Individual TD Skill Area Overlay for Scenario M-2' below).

Other resources such as Gridded Reference Graphics (GRGs) (geographical maps) were also included. Additionally, individual Skill Area IAAR Job Aids were provided to guide team self-correction during the IAAR (see Fig. 3. 'TC3 AAR Job Aid example' below). These job aids contained each learning objective (expected behavior) and questions to ask requiring squad members to monitor and reflect on their own and their squad's performance following the Force of Four framework. Finally, a Set Goals Job Aid was used to allow the squad to identify, prioritize, and set goals. So each IAAR provided opportunities to review learning objectives, discuss performance, and agree on goals.

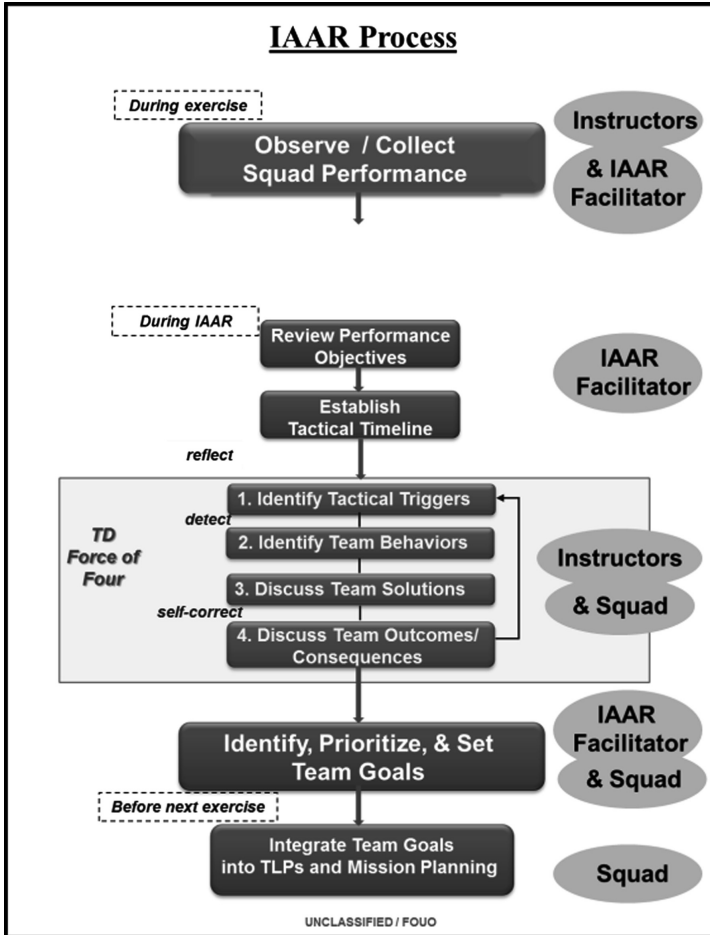


Fig. 1. IAAR Process (Source: Mitre Corporation).

In order for an IAAR to work effectively, there must be a synergy between instructors and an IAAR Facilitator. The instructors, as experts in their areas, must be allowed to contribute to the IAAR by asking specific questions and guiding discussion related to their skill areas. The IAAR Facilitator must express the tactical scenario expertise and have an overall basic understanding of the skill areas to be able to offer an integrated perspective to the squad.

An IAAR rich with learning objectives across multiple skill areas demands active participation from all squad members and skilled facilitation. Allowing squad members to contribute freely allows them to be accountable and share their perspectives of what happened, why it happened, and how to learn from the experience without reservation. The SOvM IAAR Facilitator, skill area instructors, and squad members engaged in effective IAAR questioning, feedback, and response techniques during scenario-based

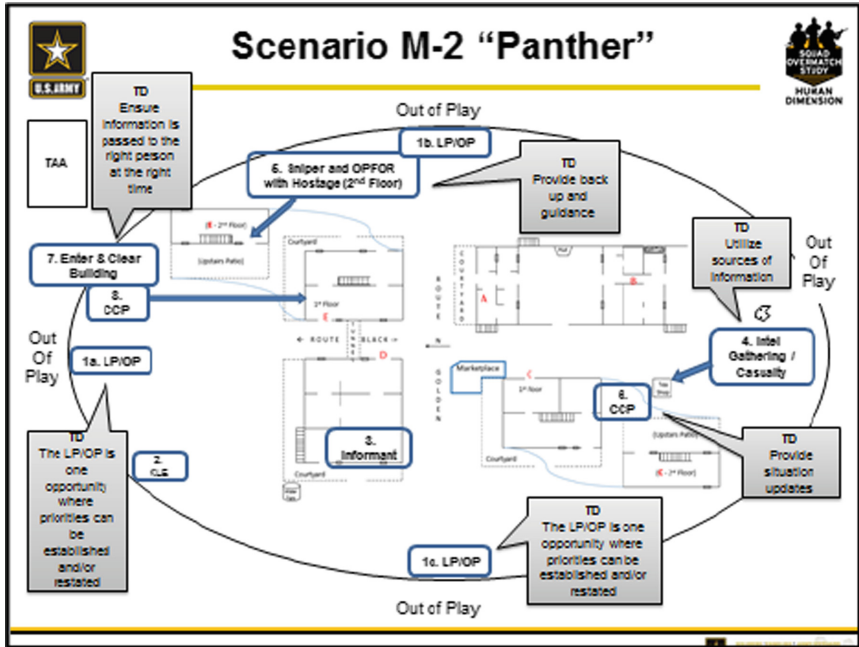


Fig. 2. Individual TD Skill Area Overlay for Scenario M-2 (Source: Mitre Corporation)

| IAAR Questions for TC3 Instructor | | |
|--|--|--|
| B-1 | | |
| What Happened? <i>Far Ambush – CUF. Two Military Casualties.</i> | | |
| What did the unit do? | Example IAAR Questions | What's a better way for next time? What might the consequences have been? |
| Team Member <input type="checkbox"/> provide MANDOWN status update | <input type="checkbox"/> Squad leader, what info was communicated to you about the casualties? <ul style="list-style-type: none"> How complete, clear and brief was the information you received? Was it enough for you to make a decision regarding medical tactical priorities? | <input type="checkbox"/> Solution: Is there a standard communication format for pushing complete, clear information to the LDR about casualties? <input type="checkbox"/> Consequence: if the SL doesn't have the right tactical medical information when coordinating a casualty response what might happen? |
| Squad Leader <input type="checkbox"/> coordinate team response to casualty | <input type="checkbox"/> How did you coordinate your unit's response to the casualties in CUF? <ul style="list-style-type: none"> what were your medical and tactical priorities? | <input type="checkbox"/> Solution: Could the team response have been organized better? <input type="checkbox"/> Consequence: What could happen if team members act independently in a way no one expects? |
| First Responder <input type="checkbox"/> provide appropriate care | <input type="checkbox"/> What treatment was provided on the "x"? <ul style="list-style-type: none"> right time? right location? what <i>resilience</i> technique did you use? | <input type="checkbox"/> Solution: Was there a safer place to provide treatment? <input type="checkbox"/> Consequence: what might have happened if the FR did/did not move off the "x" before treating? |
| Medic / Corpsman <input type="checkbox"/> return fire | <input type="checkbox"/> What did you do when you learned of the casualty? <ul style="list-style-type: none"> what <i>resilience</i> technique did you used to stay focused on the tactical priority? Any? | <input type="checkbox"/> Solution: During CUF, what is the priority for all combatants? <input type="checkbox"/> Consequence: What could have happened if the medic ran to the casualty? Did not return fire? |

Fig. 3. TC3 AAR Job Aid (one page example from Scenario B-1)

training, fostering an environment that was conducive to detection, reflection, and self-correction (see Fig. 4. ‘IAAR with soldiers at Schofield Barracks, HI’ below).



Fig. 4. IAAR with soldiers at Schofield Barracks, HI (Source: Mitre Corporation).

4 Lessons Learned

Overall, the AAR approach used in the control condition and the IAAR approach in the experimental condition was well received during the TEE. Self-report surveys revealed the majority of soldiers in both conditions rated the AAR climate following the live scenarios as strongly supportive and positive [5].

Many procedures implemented for the IAAR were successful throughout the TEE. We found that having individual instructors for each skill area instead of one instructor or only the IAAR Facilitator attempting to cover all the integrated, yet distinct, skill areas in the IAAR proved to be a good approach. Each of the SOvM skill areas were condensed from much longer program of record courses. One instructor would have had a difficult time understanding the objectives of each area, within each scenario, and know all the critical issues to address. A skilled expert handled this more effectively and efficiently. Similarly, it would have been challenging for the IAAR Facilitator to provide the tactical AAR as well as knowledgeably cover critical skill area objectives during the IAAR. Individual Observation Job Aids used during the scenarios were useful in quickly identifying skill area behaviors around trigger events within scenarios. These were then easily used to verify where the squad performed well and where they had challenges. This data also informed which questions to ask on the IAAR Job Aids and which areas to cover during the IAAR. The Set Goals Job Aid was used effectively and provided direction on identifying, prioritizing, and setting goals. The IAAR Facilitator was also a necessary and well-functioning role. This individual facilitated the entire IAAR, provided a tactical debrief, made sure instructors stayed on track, and ensured they each had opportunities to discuss objectives within their skills areas. Other aspects of the IAAR approach we found needed to be streamlined and improved.

Following is a listing of our lessons learned and how we modified the IAAR process to improve the approach. Most of these improvements were implemented during operational testing of SOvM training at Army and Marine Corp bases during 2017.

Lesson Learned 1. Combine the Scenario Event Timeline Display to Highlight a Sample of Learning Objectives. The Individual Skill Area Scenario Event Timelines made it challenging for the IAAR Facilitator to ensure specific learning objectives were discussed. Most of our resources developed specifically for the IAAR were focused on individual skill areas. Although it made it easier to focus on each skill area by addressing it separately, these individually focused job aids and resources made it more challenging to integrate the learning opportunities and present to the squads a unified training approach. The inter-relationships of these skill areas had been taught in the classroom, but this was not reinforced during the IAARs. This individual approach led to somewhat time consuming IAARs because each instructor needed time to discuss his objectives on separate job aids. To remedy this problem, the Individual Skill Area Scenario Event Timelines were combined for each scenario and these highlighted a sample of objectives instead of all objectives for each skill area in each scenario. Using this approach, objectives were still covered, but dispersed across scenarios, allowing opportunities for other skill area objectives to be presented together more fluidly. Only one Overlay was needed per scenario (see Fig. 5. ‘Updated Scenario Event Timeline’ below). Each instructor was provided time to cover their skill area objectives, but they focused on a subset and utilized one graphic to do so.

Lesson Learned 2. Design IAAR Job Aids to Include General Questions for Multiple Skill Areas for use by Each Instructor. We learned that the Individual Skill Area IAAR Job Aids resulted in a disjointed and less integrated IAAR. Therefore, we determined the IAAR Job Aids should be developed to include general questions that address multiple skill areas and can be used by each instructor for each scenario. Each instructor would simply need to determine which question(s) to ask to better meet specific learning objectives within a scenario, ensuring each had ample opportunities to engage with the squad in areas where performance deficiencies were observed. This determination could be made during the ‘huddle up’ suggested solution below. Certain scenarios might be better suited for one skill area (or more) over others. For example, in the SOvM scenarios, the earlier ones that focused on establishing a baseline were rich with opportunities for ASA behaviors to be observed and later scenarios that escalated with more TC3 events (e.g., casualties) provided numerous TC3 and TD (e.g., communication delivery, information exchange, supporting behavior, and initiative/leadership) behaviors to be exhibited. Discussing which and when certain skill area behaviors were largely utilized in the scenario and determining which of these are the most critical to debrief during the IAAR can each be better accomplished with integrated IAAR Job Aids. This would make it easier and provide greater flexibility for all instructors to more quickly determine IAAR direction and areas of emphasis.

Lesson Learned 3. Conduct a Huddle Up with the Facilitator and Instructors. We found that moving directly from a scenario into the IAAR led to confusion on sequence and timing of skill area discussions during the IAAR and which strengths and weaknesses to focus on. We determined that a Huddle Up for instructors and the IAAR Facilitator that takes place in between scenario end and IAAR start would ensure scenario

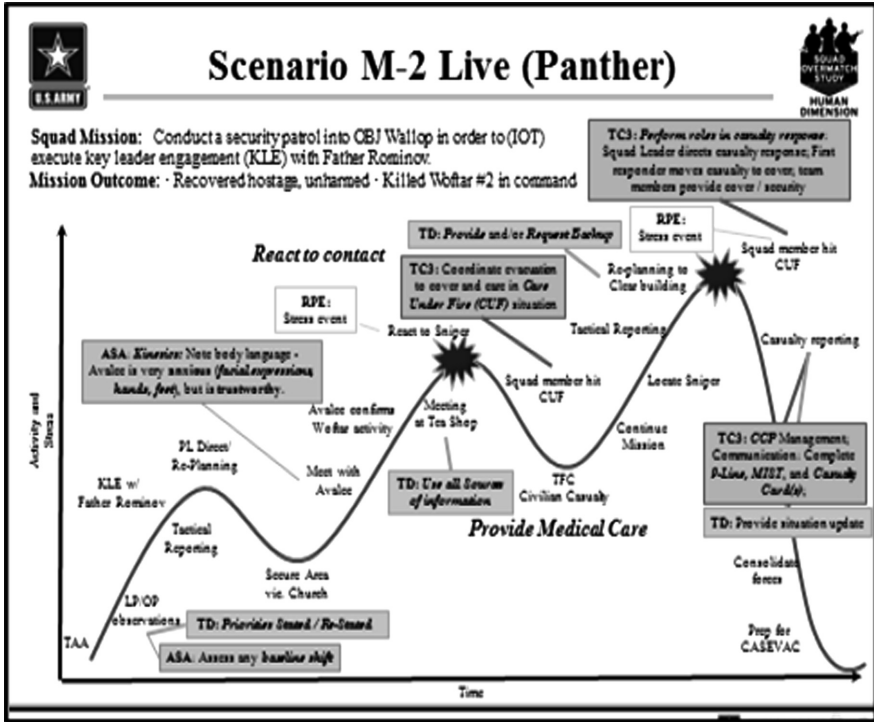


Fig. 5. Updated Scenario Event Timeline (Source: Mitre Corporation)

challenge areas were agreed upon, strengths and weaknesses within each skill area would be addressed, and provide structure and direction to the IAAR resulting in less confusion. This critical step was added to Fig. 6. ‘IAAR Process’ in between Observe/Collect Squad Performance (during exercise) and Review Performance Objectives (during IAAR). The Huddle Up is essentially a planning session for conducting the IAAR, an opportunity for instructors to exchange notes and talk with the IAAR Facilitator about which trigger events and learning objectives to focus on within key squad challenge areas and also positive aspects that should be highlighted throughout the IAAR. Preparation for the Huddle Up should take no more than 5–7 min and the same timeframe should be sufficient for the Huddle Up itself. Before the Huddle Up, instructors should review notes and tie them to events/triggers in the Scenario Event Timeline based on training objectives/performance issues. Instructors should gather performance assessment information from all sources (e.g., role players, Medic) and talk with each other about integrated learning objectives.

Huddle Up Steps are listed in Fig. 6. ‘Huddle Up Steps’ below. Accompanying video examples have been developed for SOvM operational implementation and transition efforts to emphasize key steps in the process and provide subject matter expertise in executing. During the Huddle Up, the IAAR Facilitator should ask instructors whether their training objectives were met, if there were any squad weaknesses, if goals were met, and when they want to talk in relation to the Scenario Event

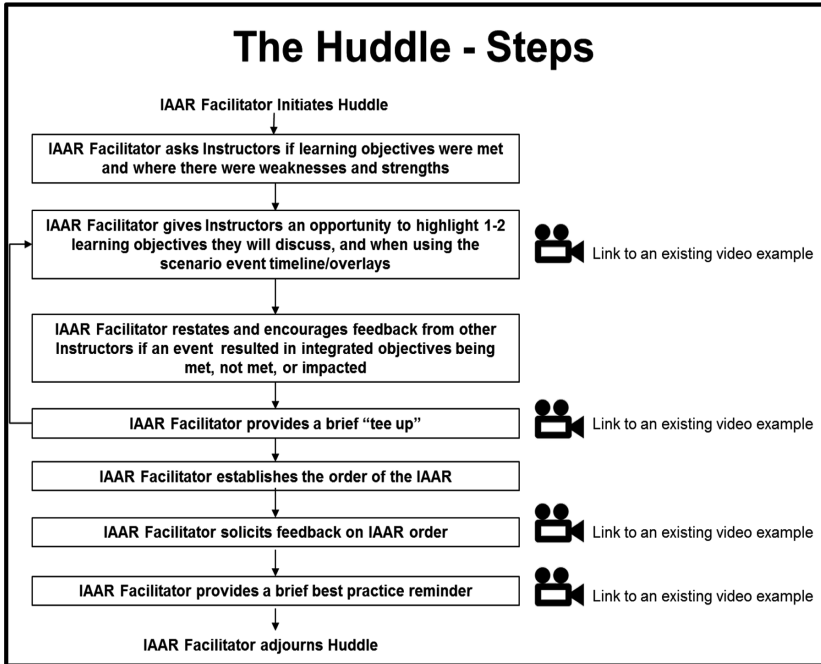


Fig. 6. Huddle Up Steps (graphic courtesy of Mitre Corporation)

Timeline during the IAAR. Each instructor should define major training objectives, related errors, and examples of good performance. The instructor and IAAR Facilitator group should determine where overall focus should be placed during the IAAR. Finally, to ensure IAAR organization and to optimally utilize the time allotted, the IAAR Facilitator should determine who will run the slide deck and take notes on goals.

5 IAAR Interactive Dashboard

Throughout the SOvM training, data are collected to support review and analysis by instructors and the IAAR Facilitator during the IAAR process. There are challenges of transforming large quantities of data into information about squad performance in time and in a form to support team development and performance improvements. One solution is the creation of a dashboard that facilitates data aggregation, synthesis, and presentation of results during the IAAR. Currently, harvesting and making sense of the data for the IAAR has been difficult to accomplish. By using a big data approach for identifying relationships among the data sources, like Observation and Assess Job Aids and IAAR Job Aids, automated field notes, GRGs with location identifiers, and audio/video, we believe sufficient, high quality performance analytics are available. An IAAR dashboard that would allow the instructors and IAAR Facilitator to enter, track and report on the Squad’s progress during each stage of training is on the drawing

board. We have conceived of an enterprise level dashboard solution that would link several data collection platforms; arrange and optimize results for reporting purposes; and deliver the information through an intuitive user interface during the IAAR. We believe that evidence-based displays not only reveal patterns of performance, they would also support near-transfer of essential feedback for team learning.

6 Conclusions

As standard as AARs have become as part of scenario-based training, IAARs provide a unique approach in ensuring multiple skill areas covered in training receive the attention needed to impact future performance. The SOvM TEE provided an environment where approaches to an IAAR could be tested and studied and valuable lessons learned derived. A number of resources are necessary for managing different skill areas and mitigating challenges with an increased number of individuals facilitating and running the IAAR. The effort involved is worth the benefits of covering multiple skill areas that complement each other when integrated and raise the potential level of learning and performance impacts. With the constrained and often limited resources available in training today, we will likely see more attempts at combining training topics and efforts. Embracing an IAAR approach can help ensure individual aspects that are combined receive similar attention to what they would have if trained separately.


Acknowledgments. The authors thank the Defense Medical Research and Development Program for sponsoring this effort. The authors also thank the Soldiers, Marines, Army Medics, and Navy Corpsmen for participating in SOvM.

References

1. Smith-Jentsch, K.A., Cannon-Bowers, J.A., Tannenbaum, S.I., Salas, E.: Guided team self-correction impacts on team mental models, processes, and effectiveness. *Small Group Res.* **39**(3), 303–327 (2008)
2. Mayer, R.E.: Aids to text comprehension. *Educ. Psychol.* **19**, 30–42 (1984)
3. Brimstin, J., Higgs, A., Wolf, R.: Stress exposure training for the dismounted squad: the human dimension. In: *The Proceedings of the Interservice/Industry Training, Simulation, and Education Conference*. NTSA, Orlando, Arlington (2015)
4. Ross, W.A., Johnston, J.H., Riddle, D., Phillips, C.H., Townsend, L., Milham, L.: Making sense of cognitive performance in small unit training. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) *AC 2016. LNCS (LNAI)*, vol. 9744, pp. 67–75. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39952-2_7
5. Johnston, J.H., Gamble, P., Patton, D., Fitzhugh, S., Townsend, L., Milham, L.: Squad overmatch for tactical combat casualty care: phase II initial findings report. In: *Program Executive Office Simulation, Training and Instrumentation*, Orlando (2016)



How Shared Screen Affected Team Collaboration Task, A Case Study of Ergonomics Experiment on Team Situation Awareness

Xu Wu^{1,2}, Chuanyan Feng², Xiaoru Wanyan², Shuang Liu³, Lin Ding¹, Chongchong Miao¹, Yuhui Wang¹, and Xueli He¹

¹ AVIC China Aero-Polytechnology Establishment, Beijing 100028, China
e126126_19@126.com

² Beihang University, Beijing 100191, China

³ Institute of Marine Technology & Economy, Beijing 100080, China

Abstract. Team situation awareness (TSA) had significant influence on collaboration work. As the essential interface of human computer interaction (HCI), visual display terminal was helpful for task performance enhancement. This study developed ergonomics experiment to investigate the impact of shared screen through team work and decision-making task. Three-member team played different roles on experiment task, and caption of the team was required to complete extra task based on others' report. During the experiment, both behavior performance and physiological measurements were recorded as indices of team situation awareness. And each team member was demanded to complete subjective rating scales to evaluate TSA afterwards. The results analysis revealed strong correlation between team situation awareness and collaborate work, while insignificant effect of shared screen usage on individual task.

Keywords: Team situation awareness · Collaboration · Shared screen
Human factor · Ergonomics

1 Introduction

Situation awareness would occur in process of individual interaction with task situation [1]. As a whole team composed of various operators, continuously observation of inside system, outside environment and team member behavior was necessary to achieve fully comprehension of current task situation, and finally formed TSA of the operator team after analysis and summary [2, 3]. Some of operation tasks could be completed by the individuals such as primary task of visual search, however, facing to rapid changing situation in the battlefield, it was difficult for single operator to understand the whole situation by constantly monitoring all update information.

In consideration of safety and function, observation and comprehension of overall situation was vital for the operator team, especially in complex system and urgent situation, where collaboration teamwork would be expected to perform tasks, including target tracking, distinguish, analysis and decision making [4–6]. According to assigned duty and task, individual operator would observe target information from human computer

interface, and achieve decision making and judgment of the operator team through communication and collaboration.

Furthermore, TSA had tremendous influence on effectiveness and efficiency of collaboration work, while both interface design and team collaboration were closely correlated with TSA [7, 8]. Therefore, friendly design of human computer interface and reasonable team collaboration would be more helpful to improve TSA, which led to insurance of efficient interaction and accurate team decision. Recently, TSA has become popular topic of SA research, and many scholars developed their studies on measurement and modeling of TSA [9]. She and Li reviewed and compared various theories of TSA in terms of definitions, conceptual models and theoretical underpinnings, and also provided major controversies on TSA for a dialectical view on the TSA theories [10]. In addition, they developed a new toolkit of digital interface to enhance mutual awareness, and explored knowledge-based tasks of team behavior and performance [11, 12]. The counter-balance could also be found that the increase in mutual awareness led to a reduction of individual situation awareness, possibly due to the limited mental resources.

In this study, to investigate whether HCI with shared screen would affect team work with improved TSA, an ergonomics experiment of typical collaboration task was carried out, and task performance, eye-movement tracking, physiological measurement, and self-rating questionnaire were used to realize multi-dimension ergonomics evaluation.

2 Method

2.1 Experiment Design

Single channel of visual task was selected in this experiment that simulated operation task of observation and comprehension of current situation. The experiment factor was designed as the usage of shared screen. And the experiment team crew was formed by three members that the captain (role c) was placed in the middle, and the other two (role a and b) were placed either side of the captain. Each member was required to interact with computer screen through normal mouse and keyboard. The experiment interface was

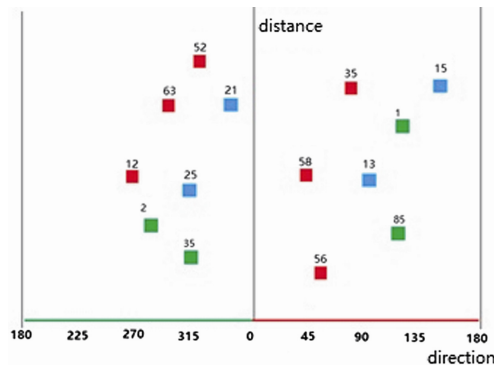


Fig. 1. Experiment interface (Color figure online)

displayed in shared (trial A) or single screen (trial B), which showed moving targets with color coding of red, green, blue from different direction and distance, as shown in Fig. 1.

2.2 Apparatus

As shown in Fig. 2, SMI Glass was adopted to measure eye-movement indices of fixation, saccade, and pupil diameter, while BioPac MP150 was adopted to measure Heart rate variance (HRV) and respiration in this experiment. And Team-Situation Awareness Rating Technology (T-SART), role conflict and ambiguity scale, and self/partner performance scale were adopted as afterwards evaluation of subjective questionnaire.



Fig. 2. SMI glass and BioPac MP150

2.3 Participants

Twenty-four graduate students from Beihang University were recruited in this experiment, with age from 22 to 27 years old and normal vision no less than 1.0. They were all informed with the detail of experiment task and procedure, and voluntarily agreed to participate in the experiment. They were divided into eight teams as three members in one team.

2.4 Experimental Task

The experiment task was designed accordingly by different roles of team members, and each one of them was required to perform three fundamental tasks, while the captain was required to perform an extra work of decision making. The specific fundamental tasks involved with direction identification task, number counting task, and memory retention task.

Direction Identification. Direction identification task required each team member to monitor visual information shown in experiment interface, including its color, direction, distance, number and moving trend. When the question box of this task was activated, the participants were required to track the mentioned target and report its direction.

Number Counting. Followed by the first task, number counting task required each team member to collect targets of certain color from all interfaces and input its total number. Meanwhile, the input results of each member were reported and shown in captain's interface, which was necessary to help the captain perform the decision making task.

Memory Retention. Memory retention task required each team member to keep real time awareness of task situation shown in the interface, and prepare to response to the randomly occurring questions of target information. The specific questions of memory retention task were involved with both individual and team SA of three levels of perception, cognition and projection.

Decision Making. The exclusive task of decision making required the captain to obtain the whole team results of the number counting task, and calculate "threat index" by the given formula, then report the current threat level accordingly.

3 Result

3.1 Behavior Performance

Direction Identification Task. Descriptive statistics analysis showed the team performance of direction identification task in trial A was better than trial B with slightly slower response time (170 ms, 3.4%) and more accurate direction deviation (0.8°, 9.7%). The average correct rate of team task was 0.928 ± 0.042 in trial A and 0.928 ± 0.040 in trial B, the average response time of team task was 5.069 ± 0.582 ms in trial A and 4.899 ± 0.568 ms in trial B. And the average direction deviation of team task was $5.642 \pm 0.973^\circ$ in trial A and $6.251 \pm 1.208^\circ$ in trial B. T-test analysis result showed significant difference between each trial in direction deviation ($T = -2.581$, $p = 0.036$) while none was found in correct rate ($T = 0.000$, $p = 1.000$) and response time ($T = 0.871$, $p = 0.412$).

The specific results of each team member were shown in Table 1. Repeated measurement two-way analysis of variance (ANOVA) was used to examine main effect on usage of shared screen (effect 1) and role of team member (effect 2). And the results showed none significant differences in main effect or interaction effect (effect 3) on both correct rate, response time and direction deviation, as shown in Table 2.

Table 1. Behavior performance of direction identification task for each team member (M ± SD)

| Team role | Correct rate | | Response time (ms) | | Direction deviation (degree) | |
|-----------|---------------|---------------|--------------------|---------------|------------------------------|---------------|
| | Trial A | Trial B | Trial A | Trial B | Trial A | Trial B |
| a | 0.913 ± 0.076 | 0.925 ± 0.038 | 5.320 ± 0.796 | 5.051 ± 0.672 | 5.740 ± 1.816 | 7.020 ± 3.539 |
| b | 0.947 ± 0.039 | 0.928 ± 0.036 | 5.009 ± 0.985 | 4.791 ± 0.822 | 5.305 ± 1.044 | 5.762 ± 1.099 |
| c | 0.925 ± 0.057 | 0.931 ± 0.061 | 4.876 ± 0.571 | 4.855 ± 0.865 | 5.881 ± 1.565 | 5.971 ± 1.003 |

Table 2. ANOVA results of direction identification task

| Effect | Correct rate | | Response time | | Direction deviation | |
|--------|-----------------|-------|-----------------|-------|---------------------|-------|
| | F-test | p | F-test | p | F-test | p |
| 1 | F(1,21) = 0.000 | 1.000 | F(1,21) = 1.559 | 0.226 | F(1,21) = 2.436 | 0.133 |
| 2 | F(2,21) = 0.474 | 0.629 | F(2,21) = 0.472 | 0.630 | F(2,21) = 0.536 | 0.593 |
| 3 | F(2,21) = 0.411 | 0.668 | F(2,21) = 0.308 | 0.738 | F(2,21) = 0.811 | 0.458 |

Number Counting Task. Descriptive statistics analysis showed the team performance of number counting task in trial A was much better than trial B with slightly higher correct rate (0.012, 1.4%) and shorter response time (1665 ms, 23.2%). The average correct rate of team task was 0.866 ± 0.067 in trial A and 0.854 ± 0.065 in trial B, the average response time of team task was 5.524 ± 0.597 ms in trial A and 7.189 ± 0.816 ms in trial B. T-test analysis result showed significant difference between each trial in response time ($T = -5.321, p = 0.001$) while none was found in correct rate ($T = 0.840, p = 0.428$).

The specific results of each team member were shown in Table 3. Repeated measurement two-way analysis of variance was also adopted. And no significant differences of correct rate was found in main effect or interaction effect, while significant differences between response time was found only in main effect on usage of shared screen, as shown in Table 4.

Table 3. Behavior performance of number counting task for each team member (M ± SD)

| Team role | Correct rate | | Response time (ms) | |
|-----------|---------------|---------------|--------------------|---------------|
| | Trial A | Trial B | Trial A | Trial B |
| a | 0.913 ± 0.076 | 0.925 ± 0.038 | 5.320 ± 0.796 | 5.051 ± 0.672 |
| b | 0.947 ± 0.039 | 0.928 ± 0.036 | 5.009 ± 0.985 | 4.791 ± 0.822 |
| c | 0.925 ± 0.057 | 0.931 ± 0.061 | 4.876 ± 0.571 | 4.855 ± 0.865 |

Table 4. ANOVA results of number counting task

| Effect | Correct rate | | Response time | |
|--------|-----------------|-------|------------------|-------|
| | F-test | p | F-test | p |
| 1 | F(1,21) = 0.506 | 0.485 | F(1,21) = 13.779 | 0.001 |
| 2 | F(2,21) = 0.413 | 0.667 | F(2,21) = 1.211 | 0.318 |
| 3 | F(2,21) = 2.136 | 0.143 | F(2,21) = 2.560 | 0.101 |

Memory Retention Task. Descriptive statistics analysis showed the team performance of memory retention task in trial A was slightly better than trial B with almost same correct rate (0.012, 2.5%) and shorter response time (557 ms, 9.5%). The average correct rate of team task was 0.384 ± 0.078 in trial A and 0.394 ± 0.053 in trial B, the average response time of team task was 6.390 ± 1.172 ms in trial A and 5.833 ± 1.385 ms in trial B. T-test analysis result showed no significant difference between each trial in response time ($T = -0.277, p = 0.790$) or correct rate ($T = 1.812, p = 0.113$).

The specific results of each team member were shown in Table 5. Repeated measurement two-way analysis of variance was also adopted. And no significant differences of correct rate was found in main effect or interaction effect, while only critical significant differences between response time was found in main effect on usage of shared screen, as shown in Table 6.

Table 5. Behavior performance of memory retention task for each team member (M ± SD)

| Team role | Correct rate | | Response time (ms) | |
|-----------|-------------------|-------------------|--------------------|-------------------|
| | Trial A | Trial B | Trial A | Trial B |
| a | 0.913 ± 0.076 | 0.925 ± 0.038 | 5.320 ± 0.796 | 5.051 ± 0.672 |
| b | 0.947 ± 0.039 | 0.928 ± 0.036 | 5.009 ± 0.985 | 4.791 ± 0.822 |
| c | 0.925 ± 0.057 | 0.931 ± 0.061 | 4.876 ± 0.571 | 4.855 ± 0.865 |

Table 6. ANOVA results of memory retention task

| Effect | Correct rate | | Response time | |
|--------|-------------------|-------|-------------------|-------|
| | F-test | p | F-test | p |
| 1 | $F(1,21) = 0.125$ | 0.727 | $F(1,21) = 3.741$ | 0.067 |
| 2 | $F(2,21) = 1.677$ | 0.211 | $F(2,21) = 0.029$ | 0.971 |
| 3 | $F(2,21) = 1.755$ | 0.197 | $F(2,21) = 1.538$ | 0.238 |

Decision Making Task. Descriptive statistics analysis showed the caption performance of decision making task in trial A was better than trial B with higher correct rate (0.075, 9.3%) and shorter response time (2487 ms, 15.0%). The average correct rate of team task was 0.878 ± 0.091 in trial A and 0.803 ± 0.113 in trial B, the average response time of team task was 14.118 ± 0.887 s in trial A and 16.605 ± 1.710 s in trial B. T-test analysis result showed significant difference between each trial in response time ($T = -3.801, p = 0.007$) but none was found in correct rate ($T = 1.871, p = 0.104$).

3.2 Eye Movement Tracking

According to the results of eye movement tracking, measurement indices were selected as fixation, saccade, blink and pupil diameter. The descriptive statistics analysis results of each trial were shown in Table 7. T-test analysis was used to examine main effect on usage of shared screen. The results showed only critical significant difference between each trial in pupil diameter ($T = -2.124, p = 0.078$) but none was found in fixation dwell time ($T = -0.512, p = 0.627$), fixation frequency ($T = -0.512, p = 0.627$), saccade

amplitude ($T = -0.670$, $p = 0.528$), saccade frequency ($T = -0.091$, $p = 0.930$), or blink rate ($T = -0.486$, $p = 0.644$).

Table 7. Eye movement tracking results ($M \pm SD$)

| Experiment index | Trial A | Trial B |
|------------------------------------|-----------------|-----------------|
| Fixation dwell time (ms) | 242 \pm 25 | 245 \pm 35 |
| Fixation frequency (times per min) | 167 \pm 15 | 168 \pm 20 |
| Saccade amplitude (degree) | 6.78 \pm 1.56 | 7.18 \pm 2.70 |
| Saccade frequency (times per min) | 145 \pm 10 | 146 \pm 21 |
| Blink rate (times per min) | 22 \pm 11 | 23 \pm 12 |
| Pupil diameter (mm) | 3.37 \pm 0.63 | 3.51 \pm 0.73 |

3.3 Physiological Measurement

According to the results of physiological measurement, experiment indices were selected as HRV and respiration. The descriptive statistics analysis results of each trial were shown in Table 8. T-test analysis was also used, and the results showed no significant difference between each trial in RR interval ($T = 0.702$, $p = 0.506$), heart rate ($T = -0.679$, $p = 0.519$), or respiration rate ($T = -0.099$, $p = 0.924$).

Table 8. Physiological measurement results ($M \pm SD$)

| Experiment index | Trial A | Trial B |
|----------------------------------|---------------|---------------|
| RR interval (ms) | 903 \pm 131 | 865 \pm 127 |
| Heart rate (times per min) | 68 \pm 11 | 71 \pm 11 |
| Respiration rate (times per min) | 16 \pm 2 | 16 \pm 2 |

3.4 Subjective Rating Scales

Subjective rating scales were selected as T-SART (Team-Situation Awareness Rating Technology), role conflict and ambiguity scale and self/partner performance scale. The T-SART results of each team member were 17.0, 21.9 and 20.3 accordingly in trial A while 16.8, 20.5 and 18.5 in trial B. Repeated measurement two-way analysis of variance showed insignificant effect on usage of shared screen ($p = 0.115$) and role of team member ($p = 0.338$).

And results of role conflict and ambiguity scale of each team member were 13.9, 13.8 and 13.5 accordingly in trial A while 12.3, 12.5 and 11.1 in trial B. Repeated measurement two-way analysis of variance showed significant effect on usage of shared screen ($p = 0.001$) but insignificant effect on role of team member ($p = 0.767$).

Moreover, self-rating results of each team member were 14.4, 12.8 and 12.4 accordingly in trial A while 13.4, 11.4 and 12.8 in trial B. Repeated measurement two-way analysis of variance showed insignificant effect on usage of shared screen ($p = 0.341$) but critical significant effect on role of team member ($p = 0.090$). In addition, partner-rating results of each team member were 14.8, 15.0 and 13.1 accordingly in trial A while 15.4, 13.9 and 14.5 in trial B. Repeated measurement

two-way analysis of variance showed insignificant effect on usage of shared screen ($p = 0.662$) and role of team member ($p = 0.559$).

4 Discussion

To examine how shared screen influenced team collaboration, task performance and physiological measurement as well as subjective rating scales were used to evaluate TSA during the experiment task. And the results of descriptive statistics analysis and repeated measured ANOVA revealed significant main effect of shared screen.

The task performance showed equally result between each trial in direction identification task, however, the shared screen could effectively accelerate the response to questions of number counting task. Therefore, shared screen was inclined to help performance enhancement of task where team collaboration was in dominant rather than individual work. In addition, the result of memory retention task was unexpected and lower than 50% in correct rate. It was mainly caused by the tremendous amount of visual information in experiment situation so that the participants were incapable of keeping short-time memory of the whole and chose reckless answers in regardless of time pressure. Moreover, shared screen had a positively effect on reduction of response time but no obviously improvement of correct rate because the time pressure was set as medium level in the experiment. So that the participants could spent more time to complete their tasks as compensation and also achieve high correct rate.

Although the results of eye movement tracking and physiological measurement showed no significant difference between each trial, the effect of shared screen had certain influence on TSA and team workload, which was proved by the critical significant effect only on pupil diameter. Since such measurement in this experiment was failed to reveal considerable interaction between physiological index and TSA, further study should concentrate on the physiological measurement and evaluation indices of TSA to implement analysis of team collaboration work.

Besides, the subjective rating scales were also used to investigate the effect of shared screen. However, T-SART and self/partner rating scales were unable to illustrate significant difference of such effect while role conflict and ambiguity scale was successfully proved to find it. And the overall results of three scales were mainly consistent with task performance. Interestingly, the third scale revealed that the self-rating point seemed to be slightly lower than that of partner-rating. The participants preferred to be strict with themselves and tolerate with others, which was mainly caused by the characteristics of culture custom and education background.

5 Conclusion

In conclusion, the experiment results showed that the task performance with shared screen was significantly more outstanding with shorter time of the overall task and higher accuracy, especially for the number counting task where team collaboration was urgent. However, there was no significant difference found in eye-movement tracking and physiological measurement between the usages of shared screen. Moreover, according

to the result of single measured index, the participants with shared screen had lower fixation rate, lower blink rate and lower saccade rate, which seemed to demonstrate the usage of shared screen was helpful to improve TSA and reduce workload to a certain extent.

Acknowledgements. The co-authors would like to thank the kindly support by Beihang University and Institute of Marine Technology & Economy.

References

1. Wei, H.Y., Zhuang, D.M., Wanyan, X.R., et al.: An experimental analysis of situation awareness for cockpit display interface evaluation based on flight simulation. *Chin. J. Aeronaut.* **26**(4), 884–889 (2013)
2. Liu, S., Wanyan, X.R., Zhuang, D.M., et al.: Situational awareness model based on attention allocation. *J. Beijing Univ. Aeronaut. Astronaut.* **40**(08), 1066–1072 (2014)
3. Liu, S., Wanyan, X.R., Zhuang, D.M.: Modeling the situation awareness by the analysis of cognitive process. *Bio-Med. Mater. Eng.* **24**(6), 2311–2318 (2014)
4. Endsley, M.R.: Situation awareness: progress and directions. In: *A Cognitive Approach to Situation Awareness: Theory, Measurement and Application*, pp. 317–341 (2004)
5. Endsley, M.R., Robertson, M.M.: Situation awareness in aircraft maintenance teams. *Int. J. Ind. Ergon.* **26**(2), 301–325 (2000)
6. Wickens, C.D.: Situation awareness: review of Mica Endsley's 1995 articles on situation awareness theory and measurement. *Hum. Factors* **50**(3), 397–403 (2008)
7. Wu, X., Wanyan, X., Zhuang, D., Liu, S.: Pilot situational awareness modeling for cockpit interface evaluation. In: Harris, D. (ed.) *EPCE 2016. LNCS (LNAI)*, vol. 9736, pp. 476–484. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_46
8. Wu, X., Feng, C., Wanyan, X., Tian, Yu., Huang, S.: Dynamic measurement of pilot situation awareness. In: Harris, D. (ed.) *EPCE 2017. LNCS (LNAI)*, vol. 10276, pp. 306–316. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58475-1_23
9. Feng, C.Y., Wanyan, X.R., Chen, H., et al.: Research on situation awareness model and its application based on multiple-resource load theory. *J. Beijing Univ. Aeronaut. Astronaut.* (2018). <https://doi.org/10.13700/j.bh.1001-5965.2017.0532>
10. She, M., Li, Z.: Team situation awareness: a review of definitions and conceptual models. In: Harris, D. (ed.) *EPCE 2017. LNCS (LNAI)*, vol. 10275, pp. 406–415. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58472-0_31
11. She, M., Li, Z.Z.: Design and evaluation of a team mutual awareness toolkit for digital interfaces of nuclear power plant context. *Int. J. Hum.-Comput. Interact.* **33**(9), 744–755 (2017)
12. Yuan, X.H., She, M.R., Li, Z.Z., et al.: Mutual awareness: enhanced by interface design and improving team performance in incident diagnosis under computerized working environment. *Int. J. Ind. Ergon.* **54**, 65–72 (2016)



Effect of Different Information Push Mechanism on Driver's Situation Awareness

Bowen Zheng, Xiaoping Jin^(✉), Zhenghe Song, Yeqing Pei,
and Xuechao Ma

China Agriculture University, Beijing 100081, China
{zbowen, jinxp}@cau.edu.cn

Abstract. The situational awareness in the field of road traffic refers to the perception, understanding and prediction of many elements in the traffic environment. Driver's situational awareness is an important factor affecting driving safety. Therefore, improving the level of situational awareness can help drivers to reduce their workload and operation errors so as to effectively protect drivers' personal safety. In this study, the subjects completed the operation by driving a vehicle in a virtual operation scenario. Through experiments, this study investigates the influence of different task complexity and different information push on the situational awareness of drivers. In the experiment, the subjects completed the task of work with information prompt and no information prompt under the two kinds of scenes (one is single, multi straight path and short distance of road condition, another is complex road condition, multiple bend, long distance). The methods of information prompting include path map prompts and voice prompts through visual display. All the subjects need to complete the following tasks. First of all in a single unfamiliar environment, the subjects observed the field work independently and completed the task, without the information hints of the task. Second subjects complete the field task with the aid of the task path planning map. Third subjects complete the task under the real-time path voice prompt. In the experiment, The subjective evaluation scoring method and Detection Response Task (DRT) are used to measure the level of situational awareness and the time and quality of the task are used to measure performance of drivers.

The results show that under the same conditions, the level of awareness of the driver under the prompt of the voice is higher than that of the driver under the forerunner of the mission path planning, and the level of situational awareness under the simple operation environment is higher than that of the complex operation environment. By analyzing the performance of work, we find that with the improvement of the level of situational awareness, the time of the task is shorter and the performance of the work becomes better.

This research is of great significance to training and improving the level of awareness of drivers.

Keywords: Situational awareness · Information push · Performance

1 Introduction

1.1 The Concept of Situational Awareness

The concept of situation awareness was first proposed and developed in the aviation field. Although SA has some similarities in the area of flight, driving definitions of SA are still needed in the area of driving to identify potentially influential Mission, environment and personal factors [1]. Situation awareness is the operator's perception of the elements in the environment within a given time and space, his or her understanding of the meaning, and the prediction of its future state of development [2].

In traffic psychology, situation awareness is the driver's perception, understanding, and prediction of many elements of the vehicle's condition, traffic signs and signals, traffic information, weather conditions, etc. throughout the road environment. In general, driving task involves five-phased of information processing function, including perception, comprehension and projection, as well as a decision on a course of action and implementing the action. The perception, comprehension and projection functions are the basis for driver situation awareness [3]. In traffic psychology, situational awareness refers to drivers' perception, understanding and prediction of vehicle status, traffic signs and signals, road information, weather conditions and other elements in the whole road environment. At present, the research on driver's situational awareness is roughly divided into three points of view. The first is from a psychological point of view, it focuses on the study of situational awareness as a simple internal cognitive phenomenon, which is mainly related to the research and analysis of drivers. The second is from an engineering point of view, it focuses on the study of situational awareness through different vehicle technology and road infrastructure. It mainly analyzes artifacts. And the third is from a systematic engineering point of view, it focuses on the study of situational awareness from the interaction of drivers, artifacts, and the interaction between both of them [4].

Three-level model of situation awareness is first proposed by Endsley. Endsley divides situational awareness into three different levels of processing (Fig. 1) [2].

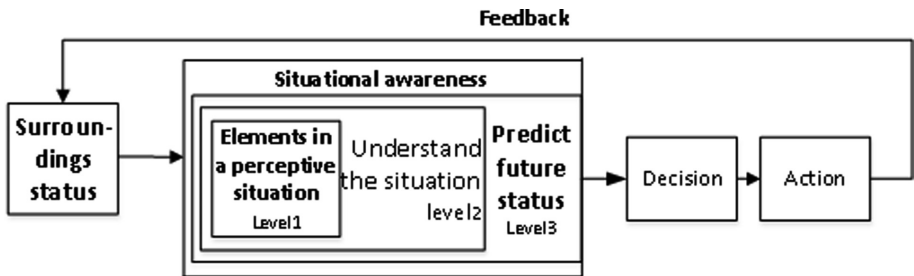


Fig. 1. Situational awareness processing

In the first level of situation awareness, the driver perceives the various factors in the surrounding environment (perception). In the second level, the driver needs to combine his past relevant experience with the key elements of the surrounding

environment, such as traffic environment and flight environment, to form a kind of coherent picture to understand (comprehend). In the third level, drivers make their own predictions about the future development and trends of these information elements (projection). Endsley's situational awareness model is a typical feedback loop model in the field of situational awareness. It is the situational awareness that feeds back the current state of the environment to the individual. Then, the situational awareness feeds back the result of integrating the current environment to the individual decision-making mechanism and makes an appropriate behavioral response. When the behavioral response can change the current unfavorable environmental state. The entire feedback loop is over.

The research by Ma investigated the effects of an adaptive cruise control (ACC) system, and cell phone use in driving, on a direct and objective measure of SA, and assessed the competition of multiple driving and communication tasks for limited mental resources in terms of driving performance. Results indicated use of the ACC system to improve driving task SA under typical driving conditions, and to reduce driver mental workload. However, the cell phone conversation caused deleterious effects on driving SA and increased driver mental load [3].

Matthews et al. proposed a contextual awareness information processing model based on the three-level model of contextual awareness. According to the different mission objectives they divided the driving behavior into three categories. The first is operational driving, its meaning is to ensure that the normal operation of their vehicles by the operation. For example, driving a car on the road to pay attention to traffic lights and speed limit signs. The second is tactical driving, it refers to what the driver does to ensure safe interaction with other vehicles. Such as driving cars on the road to overtake and change lanes. The third category is strategic driving, it refers to the driver of the known information of high-level reasoning, calculation process of thinking about driving strategy. For example, acceleration is slowed down in advance. We can see that operational driving relies on the first level of the three-level model, tactical driving relies on the second and third levels, and strategic driving relies on all three levels [5].

Salmon et al. found that A high level of situational awareness does not necessarily depend on a low level of situational awareness. Some very experienced drivers can skip a low level of situational awareness directly to get a higher level of situational awareness [6].

Walker et al. found that an approach to SA based on Neisser's perceptual cycle theory is anchored to a network based methodology. It is applied within the context of a longitudinal on-road study involving three groups of 25 drivers, all of whom were measured pre- and post-intervention. One experimental group was subject to advanced driver training and two further groups provided control for time and for being accompanied whilst driving. Empirical support is found for all five hypotheses. Advanced driving does improve driver SA but not necessarily in the way that existing situation focused, closed loop models of the concept might predict [7].

Matthews et al. outlined multiple elements of awareness defining SA in driving, including spatial awareness, identity awareness, temporal awareness, goal awareness and system awareness. They said spatial awareness refers to an appreciation of the location of all relevant features of the environment. Identity awareness refers to the knowledge of salient items in the driving environment. Temporal awareness refers to

knowledge of the changing spatial “picture” over time. Goal awareness refers to the driver's intention of navigation to the destination, and the maintenance of speed and direction. System awareness refers to relevant information on the vehicle within the driving environment, which may also be viewed as a system [5].

Gugerty and Tirre presented a similar concept of driver situation awareness. They said drivers must maintain navigation knowledge, local scene comprehension (knowledge of nearby traffic for maneuvering), knowledge of spatial orientation, and knowledge of their vehicle's status to maintain good SA during driving [8].

Gugerty and Tirre and Matthews et al. considered in-vehicle system interaction knowledge to be important in a driving environment, for example, when a car traveling at a constant speed under cruise control enters a higher speed limit area, driver awareness of their vehicle speed, the speed limit and knowledge of how to set a higher speed represents good SA [9].

Crundall and Underwood found that the differences between novices and experienced drivers in their distribution of visual attention under different levels of cognitive load imposed by different types of road, and as reflected in their visual search strategies. The results suggested that experienced drivers select visual strategies according to the complexity of the roadway, and that the strategies of novices are too inflexible to meet changing demands [10].

1.2 The Main Method of Measuring Situational Awareness

The current main method of measuring situational awareness is divided into situation awareness global assessment technique, eye movements, and Proposition networks. It features a situational awareness-inducing in a driving simulator. In the driving simulator, testers freeze the simulated tasks of the subjects at random time points, and present blank screens, allowing subjects to answer some questions referred to the task according to their memory or to make a detailed description of a specific situation. eye movements reflect the driver's traffic information for different degrees of processing through the eye movement indicators. This is the most commonly used measurement method for studying driver situational awareness. Eye movement measurement refine the driver's cognitive process, to more accurately examine the difference between different drivers situational awareness, such as experienced drivers and novices between the level of awareness of situational awareness and support for driver's situational awareness training. Proposition networks refers to use specialized software to analysis and process verbal reports of drivers and cognitive tasks after the interview, and then form a network model that contains a variety of informational elements. We describe the driver's situational awareness through the development and changes of this network model. In Proposition networks, We use the ellipse to represent the information elements in the environment, and the labeled arrows indicate the relationship between the elements [11].

2 Method

Through indoor simulated driving, DRT, timer and subjective evaluation table are used to collect data. The static experimental scene uses the PC-side driving simulation software to simulate the driving scene. It is divided into Fig. 2 (single road, multiple straights, short distance) and Fig. 3 (complicated road, multiple curves, long distance).



Fig. 2. Single road, multiple, straights, short distance



Fig. 3. Complicated road, multiple curves, long distance

11 subjects were selected, all 21–23 year-old college students, have C1 driver's license and can skillfully drive a car. First of all, we confirm their personal information, driving experience and driving habits, to ensure that the subjects were healthy and no discernible hearing disorders. Before the experiment, the training of all subjects were consistent, so that subjects were familiar with the experimental background in the same grade. The training process is to first explain to the subjects experimental principle, experimental process, experimental purpose, experimental interface, operation method. Then subjects were simulated driving, familiar with the driving environment, to meet the experimental requirements (see Fig. 4).



Fig. 4. The training process

In addition, two helpers were required for each experiment, one person was in charge of control and data recording of the DRT measurement equipment, and the other was responsible for voice prompts the subjects during the experiment. The experiment takes two experimental variables: different task complexity and different information push on the situational awareness of drivers.

Information push mode is divided into three types: First of all in a single unfamiliar environment, the subjects observed the field work independently and completed the task, without the information hints of the task. Second subjects complete the field task with the aid of the task path planning map. Third subjects complete the task under the real-time path voice prompt.

This experiment is a two-level, two-level and a three-level ($2 * 3$) mixed experiment, a total of six experiments, as shown in the Table 1.

Table 1. Experimental design table

| Experiment number | Mission complexity | Information push method |
|-------------------|--------------------|-------------------------|
| 10 | 1 (Low level) | 0 (Traffic observation) |
| 20 | 2 (High level) | 0 (Traffic observation) |
| 11 | 1 (Low level) | 1 (Map tips) |
| 21 | 2 (High level) | 1 (Map tips) |
| 12 | 1 (Low level) | 2 (Voice prompts) |
| 22 | 2 (High level) | 2 (Voice prompts) |

We numbered the experiment in order to facilitate the recording. The first digit of the experiment number represents the level of task complexity, 1 for low level and 2 for high level. The second digit represents how information is pushed, 0 for traffic observation, 1 for the map tips, 2 for voice prompts. In order to prevent the experimental order may

give the subjects the learning effect, we require different subjects have different experimental sequences. In other words, we randomized the experimental sequences, and the same scene will not be tested twice in succession.

The dependent variables in the experiment include subjective evaluation of situational awareness, DRT performance and driving performance. During the experiment, the experiment process was recorded by the camera to analyze after the experiment.

2.1 Subjective Evaluation and Analysis of Situational Awareness

After the experiment, we explained the meaning of situational awareness to the subjects. When the subjects got a good understanding of the subject, we asked the subjects to score the scores according to Table 2.

Table 2. Subjective rating

| Score | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---------|-----------|------|-----------|---------|----------------|-----------|----------------|
| Meaning | Very easy | Easy | Some easy | Neutral | Some difficult | Difficult | Very difficult |

2.2 DRT Performance

Obtain the driver's reaction ability that the reaction time and the target hit error rate through the DRT performance.

2.3 Driving Performance

Get the driving performance through the driver to complete the task of driving the situation, driving performance is divided into driving task completion time and driving the number of collisions.

3 Results and Discussion

The results obtained by the above experiment, the analysis method used is EXCEL scatter plot and histogram roughly analyzed, and then we judge the significance of the experimental results by the hypothesis t-test.

3.1 Subjective Evaluation and Analysis of Situational Awareness

After the experiment, we explained the meaning of situational awareness to the subjects. When the subjects got a good understanding of the subject, we asked the subjects to score the scores according to Fig. 5.

We averaged the subjective evaluation scores for each type of task, as we can see from the results, when no prompt was provided and the subjects only relied on the road to obtain information, the subjects rated the scores as 4.7 and 5.8, they were neutral to difficult. When there were map prompts, the scores dropped to 3.4 and 4.5, which was slightly easier than without information push, but not obvious. However, when there was

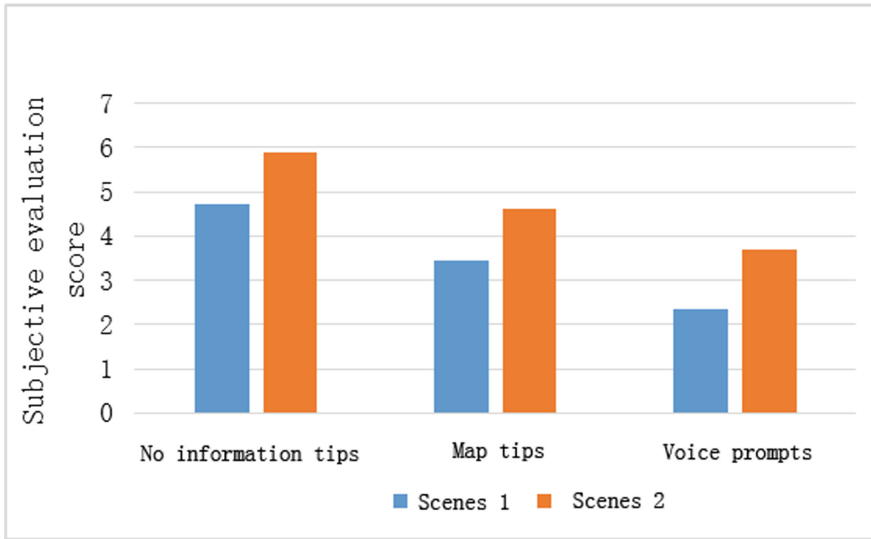


Fig. 5. Subjective evaluation

a voice prompt, the evaluation score reached 2.4 and 3.7, significantly reduced. It proves that when the other conditions are the same, the voice prompts can make the subjects better perceive the traffic information. Because subjects are required to divert part of the visual resources to get information from the map prompts and need to locate the vehicle after obtaining the information to determine their location before making decisions. When they use the voice prompts, visual resources will not be occupied, just get the voice information, compared with the current environment, they can make the appropriate decisions. Without any information push, the subjects need to observe the road conditions, pick out the information they need from a lot of information, and then perceive, understand and predict, so in this case, the subject's situation awareness rating will be high, this indicate a high level of situational awareness will make the driver's subjective evaluation score lower. By t test, all $P \ll 0.05$. Therefore, the judgment of subjective evaluation is very significant, it has statistical significance (Tables 3, 4, 5 and 6).

Table 3. Subjective evaluation t-test mean

| Analysis (Scenario 1) | | |
|---------------------------|---------------------|----------|
| | No information push | Map tips |
| Mean | 4.7273 | 3.4545 |
| Standard deviation | 0.6466 | 0.6742 |
| Observations | 11 | 11 |
| Correlation coefficient | 0.757 | |
| Assume average difference | 0 | |
| Df | 10 | |
| T | 9.037 | |
| Sig (both sides) | 3.9882E-6 | |

Table 4. Subjective evaluation t-test mean

| Analysis (Scenario 1) | | |
|---------------------------|---------------------|---------------|
| | No information push | Voice prompts |
| Mean | 4.7273 | 2.3636 |
| Standard deviation | 0.6466 | 0.6742 |
| Observations | 11 | 11 |
| Correlation coefficient | 0.480 | |
| Assume average difference | 0 | |
| Df | 10 | |
| t | 11.628 | |
| Sig (both sides) | 3.9275E-7 | |

Table 5. Subjective evaluation t-test mean

| Analysis (Scenario 2) | | |
|---------------------------|---------------------|----------|
| | No information push | Map tips |
| Mean | 5.8182 | 4.5455 |
| Standard deviation | 0.7507 | 0.6875 |
| Observations | 11 | 11 |
| Correlation coefficient | 0.793 | |
| Assume average difference | 0 | |
| Df | 10 | |
| T | 9.037 | |
| Sig (both sides) | 3.9882E-6 | |

Table 6. Subjective evaluation t-test mean

| Analysis (Scenario 2) | | |
|---------------------------|---------------------|---------------|
| | No information push | Voice prompts |
| Mean | 5.8182 | 3.7273 |
| Standard deviation | 0.7507 | 1.009 |
| Observations | 11 | 11 |
| Correlation coefficient | 0.456 | |
| Assume average difference | 0 | |
| Df | 10 | |
| t | 7.347 | |
| Sig (both sides) | 2.4621E-5 | |

3.2 DRT Performance Analysis

The DRT reaction time is the time difference between the moment when the driver presses the reaction button and the diode starts to emit light after the driver sees the light-emitting stimulus. Response time under different message prompts, all of which

are average response times. We use the DRT reaction time as the vertical axis, subjective evaluation of situational awareness as abscissa, made of a scatter plot (Figs. 6 and 7). All of response time are average response time.

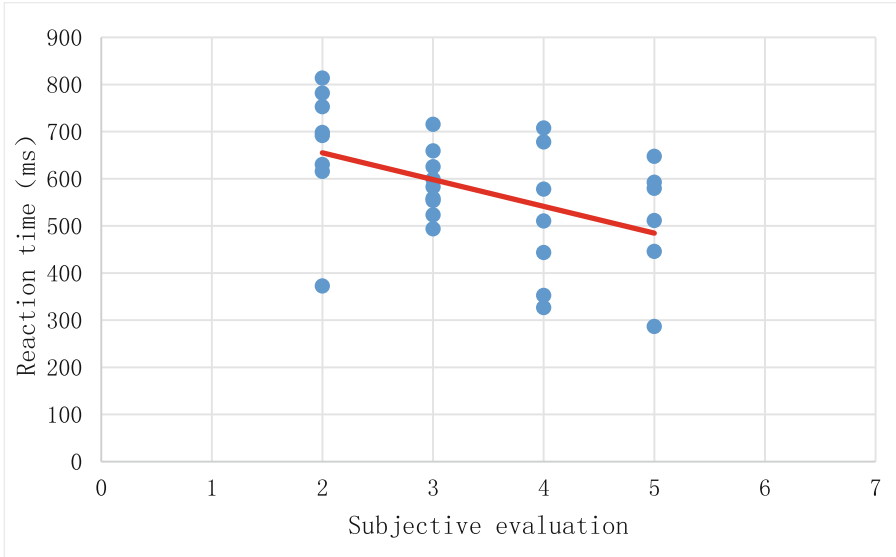


Fig. 6. Subjective evaluation - DRT scatter plot in scene 1

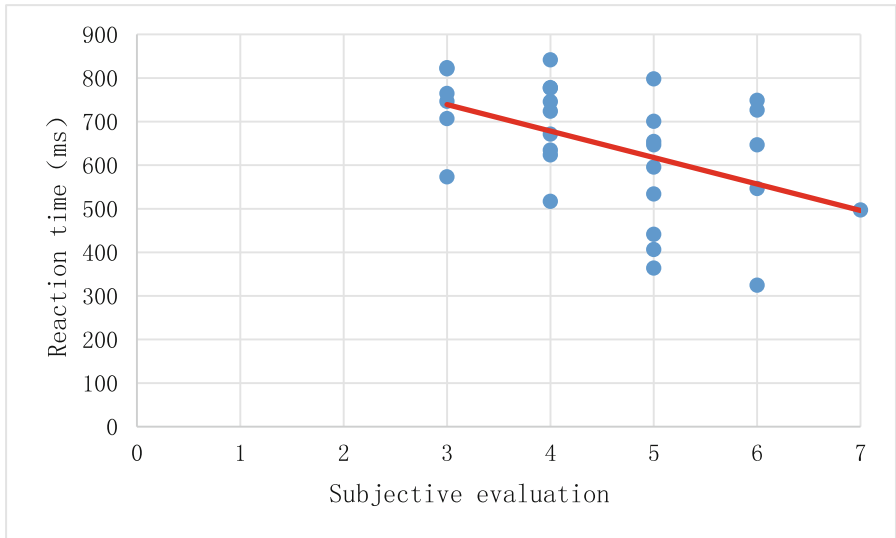


Fig. 7. Subjective evaluation - DRT scatter plot in scene 2

From the figure we can see that with the increase of situational awareness scores, the response time of subjects showed a downward trend. The results showed that in the driving process, when prompted, people’s response time will be longer. Without information push, the response time of the subjects in both scenarios was 534.29 ms and 610.26 ms. The reaction time at the map prompt is 596.11 ms and 665.83 ms and the response time at voice prompts was 667.31 ms and 717.86 ms. The reason for the longer reaction time is that on the one hand, map prompts and voice prompts can distract users thus prolonging the reaction time, on the other hand, observing the scatterplot of situational awareness and response time, we can clearly see that with the increase of situational awareness scores, subjects’ response time is reduced. Because in a strange scene, if the subject is too difficult to obtain situational information, which level of situational awareness is not enough, he can not be better integrated into his driving work. So he has enough energy on the DRT mission. As the level of situational awareness improved, subjects were able to access information easily, so he could dedicate himself to driving, his energy spent on DRT tasks became less and his reaction time lengthened. This shows that a high level of situational awareness leads to longer driver DRT response time.

3.3 Driving Completed Performance Analysis

During driving, the driver is not aware of the surrounding information (lack of situational awareness) or is wrongly aware of the surrounding information (false situational awareness), resulting in the occurrence of a collision. We put each crash as a mistake, for performance evaluation (see Fig. 8).

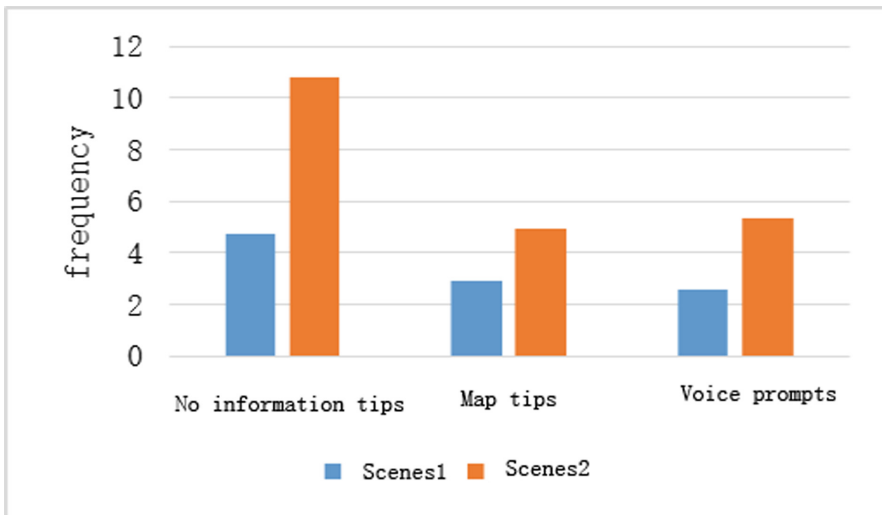


Fig. 8. Frequency

The results show that information prompts will make the driver effectively avoid collision during driving. Performance is 4.72 times and 10.82 times without any information push. 2.91 times and 4.91 times for map prompts, 2.55 times and 5.36 times for voice prompts. So under the information push, subjects were able to obtain timely and accurate information on the road, make timely adjustments to reduce the collision. By observing the video, the number of collisions that occurred at the corners under voice prompts was more than the number of prompts at the map. Because the subjects passed the curve, they were prompted by the map to have a clear understanding of the curve. However, they can only get the message that the next corner will be through voice prompts, and there is no intuitive understanding of the corner in the brain. In Scene 2, there are many corners, so the performance of the map prompts is better than the voice prompts. At the same time, we found through observation that in sharp turns, the driving performance of the voice prompt is better than the driving performance of the map. This is due to that voice prompts can convey information to the driver promptly and clearly before turning, but map information acquisition depends on whether the driver has checked the map at the turning point. Therefore, for the continuous complex road conditions, the map prompts can optimize the driver's situational awareness. For some sudden and other unexpected traffic conditions, under the voice prompts, the driver's situational awareness is better. For the continuous complex road conditions, the driver can be prompted by voice to observe the map, and at the same time the map prompts, so as to enhance the driver's situational awareness.

4 Conclusion

The results show that under the same conditions, the level of awareness of the driver under the prompt of the voice is higher than that of the driver under the forerunner of the mission path planning, and the level of situational awareness under the simple operation environment is higher than that of the complex operation environment. By analyzing the performance of work, we find that with the improvement of the level of situational awareness, the time of the task is shorter and the performance of the work becomes better.

This research is of great significance to training and improving the level of awareness of drivers.

References

1. Billings, C.E.: *Aviation Automation: The Search for A Human-Centered Approach*, vol. 41, no. 4, p. 560. Lawrence Erlbaum Associates (1996)
2. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. *Hum. Factors J. Hum. Factors Ergon. Soc.* **37**(1), 32–64 (1995)
3. Ma, R., Kaber, D.B.: Situation awareness and workload in driving while using adaptive cruise control and a cell phone. *Int. J. Ind. Ergon.* **35**(10), 939–953 (2005)
4. Stanton, N.A., Salmon, P.M., Walker, G.H., et al.: Is situation awareness all in the mind. *Theor. Issues Ergon. Sci.* **11**(1–2), 29–40 (2010)

5. Matthews, M., Bryant, D., Webb, R., et al.: Model for situation awareness and driving: application to analysis and research for intelligent transportation systems. *Transp. Res. Rec. J. Transp. Res. Board* **1779**(1), 26–32 (2001)
6. Salmon, P.M., Stanton, N.A., Walker, G.H., et al.: Measuring situation awareness in complex systems: comparison of measures study. *Int. J. Ind. Ergon.* **39**(3), 490–500 (2009)
7. Walker, G.H., Stanton, N.A., Kazi, T.A., et al.: Does advanced driver training improve situational awareness? *Appl. Ergon.* **40**(4), 678–687 (2009)
8. Gutzwiller, R.S.: Individual differences in working memory affect situation awareness. *Dissertations & Theses - Gradworks* (2011)
9. Gugerty, L., Rando, C., Rakauskas, M., et al.: Differences in remote versus in-person communications while performing a driving task. In: *Human Factors and Ergonomics Society Meeting*. SAGE Publications (2003)
10. Crundall, D.E., Underwood, G.: Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics* **41**(4), 448–458 (1998)
11. Salmon, P.M., Stanton, N.A., Walker, G.H., et al.: *Distributed Situation Awareness: Theory, Measurement and Application to Teamwork* (2009)

Psychophysiological Measures and Assessment



Mental Workload Estimation from EEG Signals Using Machine Learning Algorithms

Baljeet Singh Cheema¹, Shabnam Samima², Monalisa Sarma²,
and Debasis Samanta³(✉)

¹ Directorate General of Information System, Integrated Headquarters,
Ministry of Defence (Army) Government of India, Kharagpur, India

balema.155h@gov.in

² Subir Chowdhury School of Quality and Reliability,
Indian Institute of Technology Kharagpur, Kharagpur, India

{shabnam.samima,monalisa}@iitkgp.ac.in

³ Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur, Kharagpur, India

dsamanta@sit.iitkgp.ernet.in

Abstract. Multitasking conditions prevalent in many environments such as critical operations in defense activities, evaluating user interfaces in man-machine interaction, etc. require assessment of mental workload of operators. However, mental workload (MWL) cannot be perceived directly as it is a complex and abstract property of human physiology. The techniques available in the literature for its assessment usually depend on *subjective analysis*, *performance analysis* and *psycho-physiological measurements*. But, these approaches despite being followed often prove to be inadequate due to high inter-personal variations and inconvenient procedures and subjective to the bias of evaluators. With the recent advancements of proliferation of Brain Computer Interface (BCI) devices and machine learning algorithms, it is possible to estimate MWL automatically. Nevertheless, there is a need to address issue like managing a high dimension and high volume data in real time. In this work, we propose an approach to estimate the mental workload using electroencephalogram (EEG) signals of an operator while in operation. A thorough investigation of different features, optimization of features and selecting an optimal number of channels are the some of the crucial steps have been addressed in this work. We propose a novel feature engineering method to extract a reduced set of features and utilize only a sub-set of channels for the purpose of classification of workload into different levels with the help of supervised machine learning techniques. Further, we investigate the performance of different classifiers and compare their results. It can be inferred from the observed results that mental workload estimation using machine learning algorithms is a better solution compared to the existing approaches.

Keywords: Electroencephalography
Psycho-physiological measurement · Brain-computer interface
Mental workload estimation · Machine learning algorithms
 n -back task · Dual n -back task

1 Introduction

Rapidly increasing growth and development in various industrial sectors like aviation, transport, military or space requires multitasking and continuous vigilance from operators to perform various jobs. This often over burdens the operators by placing huge mental workload upon them and leads to work-related stress and possibility of human errors. According to [7], Mental workload (MWL) refers to the amount of resources needed for processing of a certain task. It depends on characteristics of the task, the situation and the person. It is an abstract property of human-machine interaction which is not directly observable as there exists inherent difficulties in defining MWL and in understanding the factors which describe it in the best possible manner. It also poses difficulty in building a general/robust model for predicting performance. However, in the literature the level of workload has been inferred through three prime approaches, namely (1) subjective measures, (2) performance-based measures and (3) physiological measures [7]. Subjective approaches rely on the self assessment from the subjects about the difficulty of various tasks; performance-based measures depend on user performance for determining and assessing the cognitive state [12] and physiological methods attempt to interpret the cognitive workload with the help of invasive, semi-invasive and non-invasive physiological techniques.

Out of the above-mentioned categories, physiological measurements are comparatively better as they provide continuous and objective measurement of operator state. These measurements attempt to interpret the psychological processes through their effect on the body state, rather than through task performance or perceptual ratings. There are a number of diverse techniques available in the literature under this category [8]; however, each one of them is associated with some merits and demerits. In this regard, measurement technique such as ECoG (electrocorticography) provides better spatial and temporal resolution and better signal quality. But, it is a semi-invasive technique, which requires risky surgery. On the other hand, MEG (magnetoencephalography) is a non-invasive measurement technique, but incurs huge equipment cost and is not suited for everyday applications. fNRIS (functional magnetic resonance imaging) is relatively inexpensive and portable, but provides shallow spatial resolution of the order of few centimeters, while the time resolution of around 200 ms. An EEG (electroencephalography) based mental workload assessment which in earlier days utilized costly, wired and bulky devices posed serious limitations for application in real world applications. However, recent developments in brain-computer interfaces targeting real-life applications include *wireless EEG acquisition systems* that a person can easily wear while performing everyday activities. Of late, such a low cost, portable and wireless EEG devices have gained immense popularity for

studying cognitive workload [19] and vigilance task [17], as they allow for direct mental state assessment and because of their high temporal resolution, which is in the order of milliseconds. This makes EEG an appropriate tool for capturing fast and dynamically changing brain wave patterns in complex cognitive tasks. Besides, it seems that the use of wireless data acquisition systems to assess mental workload can enable more novel applications of mental workload measurement. This development supports exploring the feasibility of wireless data acquisition devices in MWL assessment [1,2]. Therefore, in this work we aim to:

- Explore the feasibility of wireless data acquisition devices in MWL assessment.
- Estimate MWL induced via various n -back and Dual n -back tasks and extract desired features using feature engineering.
- Study the effect of channel selection and feature optimization on classification performance.
- Investigate the capability of supervised machine learning algorithms for efficient classification of mental workload.
- Study the performance accuracy obtained using item-class classification.

In this work, we have used *Emotiv Epoc+* device to explore the feasibility of inexpensive EEG devices for assessing MWL and for collecting data. We performed a variety of n -back and Dual n -back tasks for inducing different levels of mental load on participants brain. Moreover, we utilized feature engineering step for extracting and selecting most effective features. Next, for classifying the MWL we have resorted to machine learning as it is a popular research field devoted to the development of inductive models, algorithms and procedures that can learn from data, extract trends and make predictions. The choice of machine learning algorithms is done from a pool of such algorithms, so as to have an optimal configuration of these algorithms. We have chosen seven different types of machine learning approaches, namely:

- Similarity based: K -nearest Neighbours
- Information based:
 - Random Forest
 - Decision Tree
- Error based:
 - Support Vector Machines
 - * Linear
 - * Radial Basis Function (RBF) Kernel
 - Multi Layer Perceptron
- Statistics based: Linear Discriminant Analysis

The rest of the paper is organized as follows: Sect. 2 presents the literature survey of works related with the classification of mental workload. Section 3 describes materials and methodology. Section 4 elaborates the process of EEG signal analysis. Next, we discuss about the obtained results in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Literature Survey

In [5], authors estimated the mental workload, by using EEG features, for designing the intelligent learning systems. The developed workload index uses a Gaussian Process Regression model for predicting the workload of the individuals. The potential of both fNIRS and EEG (in combination) for classification of users' mental workload has been explored by the authors in [8]. In the recent years, rigorous efforts are being made to classify the mental workload into different levels. For example, in [14] classification of workload has been done using EEG based features. In [12], stepwise regression and multi-class linear classification has been utilized to extract statistical EEG features and to classify the workload into four levels. Authors in [20] have classified workload in seven levels by applying discrete wavelet transform and using artificial neural network (ANN). Further, in [9] EEG features that are sensitive enough to detect workload changes were identified. Variation of workload in different tasks has been found to be correlated with the EEG patterns in [4]. In [11], authors utilized cross-task performance based feature selection and regression model to classify mental workload. Binary classification of mental workload has successfully been done with the help of Fisher LDA and ERP based EEG features in [16]. Besides, the traditional mental workload assessment techniques were compared against the classification models built using machine learning approaches in [13].

3 Materials and Methodology

3.1 Subjects

Five healthy male and five healthy female volunteers participated in the experiment. The participants were between 20 and 24 years old, and except one, all were right handed. The participants were under-graduate and post-graduate students studying at the Indian Institute of Technology, Kharagpur. The participants had normal or corrected-to-normal vision. Further, participants were not on any medication and had no psychiatric or neurological disorders. Informed consent was taken from each participant before beginning of the experiment and were given liberty to select a time for the experiment in which they would feel *alert*. Moreover, the participants were also instructed to refrain from ingesting alcohol and/or sedative medications 24 h prior to the experiment and from caffeine and/or nicotine two hours prior to the experiment.

3.2 Data Acquisition and Experiment Protocol

The data collection has been carried out using the bluetooth enabled *Emotiv EPOC+* EEG device, having sampling rate of 128 Hz. The device comprises of 14 channels, namely AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4 plus two references (P3/P4) and follows the international 10–20 standard for locations of electrodes. A minimum distance of five meters was maintained

between all power sources and the place of experiment. Further, the use of mobile phones was prohibited inside the laboratory during the experimentation. Two dedicated systems were used for the purpose of data collection, one for data recording and the other one for running the workload generating tasks. The tasks were run on a 21" all-in-one PC kept at a distance of 75 cm from the subjects. To minimize the artifacts originating from muscular movement, that is, due to electromyographic (EMG) activity, subjects were asked to avoid unnecessary physical movements during data recording. Furthermore, their hands were placed in a fixed position such that they could easily tap their fingers on the keyboard in response to the correct answer.

3.3 Workload Generating Task

We used the open source application, namely "Brain Workshop" [10] for generating the mental workload (MWL). The n -back task available in this application is a cognitive task which is mostly used as an assessment tool in cognitive neuroscience. The main advantage of n -back task is that it does not introduce any bias due to experience of an individual participating in the experiment. In other words, repetitive experiments with same participant introduces seldom bias. In this work, we have considered two variants of n -back task (n -back and dual n -back task) for MWL generation. The n -back/dual n -back tasks have all three ingredients of cognitive load namely:

- *Intrinsic load*, which is the load induced by the inherent nature of the task being processed. The inherent difficulty of task can be increased firstly by increasing the value of 'n' from 1 to 2 and secondly by migrating from n -back to dual n -back task.
- *Extraneous load*, which is induced by external factors like time pressure, noise, situation, work organization, etc. This type of load can be increased by reducing the time between the stimuli. In our experimental setup, we kept it constant at 3 s.
- *Germane load*, which is the load placed on working memory during schema formation and automation. Such a kind of load can be increased by increasing the value of 'n' which leads to increased amount of information required to be stored and processed in the working memory.

Further, we used five different tasks to generate five different load levels, namely idle, 1-back, 2-back, dual 1-back and dual 2-back in our experimentation. During the idle task, the participants were asked to remain still with eyes closed. In the 1-back and 2-back task scenarios, a 3×3 grid was shown with stimuli appearing randomly at one of the grid locations on the screen. On the appearance of a stimuli, or trial, the participants were asked to respond whether or not the current stimulus is the same as the one that they saw n (that is, 1 or 2) presentations ago. Hence, for each trial, participants needed to memorize the previous n sequence of stimuli and perform a matching task mentally. Successfully, the dual n -back task involves remembering a sequence of spoken alphabet

and a sequence of positions of the stimuli at the same time, and identifying when an alphabet or position matches the one that appeared n trials back. Each task in the experiment had a total of 60 audio/visual stimuli (depending on various tasks) appearing after every three seconds.

3.4 Procedure

The experiments for data collection were conducted in an electrically isolated BCI laboratory under controlled environmental conditions so as to ensure adequate comfort to the participants. Here, we have performed experiments and tried to develop a method to classify mental workload not only when training and testing is done on the same task, but also when training and testing is done on different tasks. We utilized five distinct task levels in this experiment. Each participant performed all five levels of experiment successively.

Before beginning the experiment, each participant first filled the consent form and personal details form containing information about their age, gender, sleep duration, medication, status of mental health, education background, etc. Next, the experiment was started with the minimum load task, that is, idle task which was followed by the 1-back, 2-back, dual 1-back and dual 2-back tasks, respectively. In the n -back ($n = 1$ or 2) task, the participants responded to the ‘position matching’ of the stimuli by pressing the alphabet ‘A’ from the keyboard if the position of the current stimulus matched to the position of stimulus presented n -trials back. While in the dual n -back ($n = 1$ or 2), the participant pressed ‘A’ key for ‘position match’ and ‘L’ for ‘sound match’, respectively (refer Figs. 1 and 2). Switching from one task level to other was marked by a rest period of one minute. In each task level, a total of 60 trials/stimuli were presented, wherein each one appeared after every three seconds. EEG data recording for every load level of n -back task has been done for three minutes. Thus in total, for all levels, the duration of experiment was 20 min. The entire experiment protocol is graphically shown in Fig. 3.

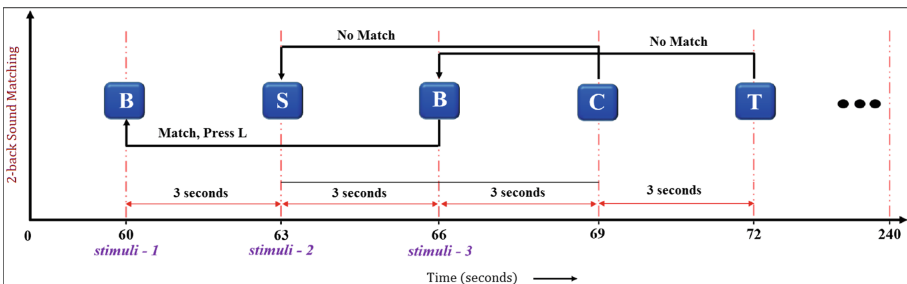


Fig. 1. An illustration to represent a 2-back task

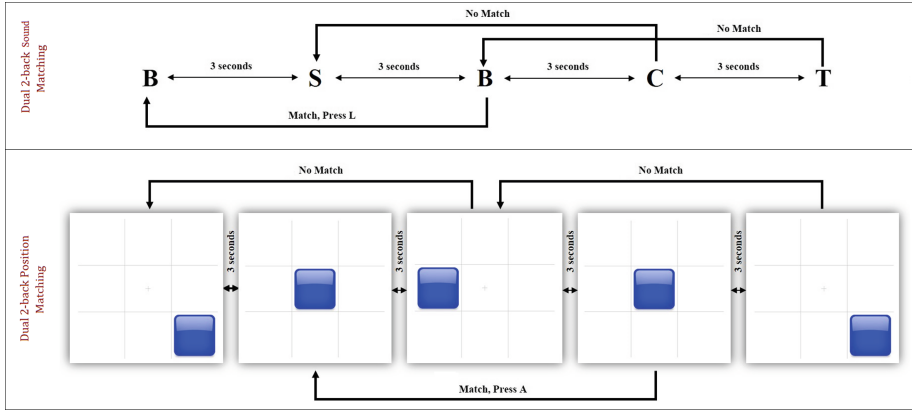


Fig. 2. An illustration to represent a dual 2-back task

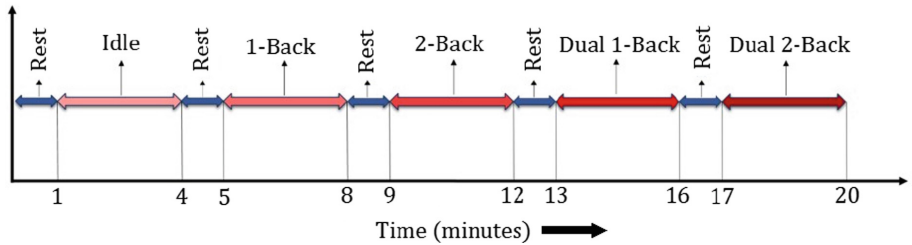


Fig. 3. Complete protocol

4 EEG Signal Analysis

The raw EEG signals captured through the scalp are contaminated with electrical signals and other undesired cerebral activities, which makes them unsuitable for feature extraction. These artifacts cause changes in the EEG measurements and severely degrade the useful signal of interest. Thus, it is necessary to process the EEG signals before we extract features from it. To process EEG signals, we begin with signal pre-processing phase which is followed by channel selection and feature extraction. Finally, we classify the data using the machine learning algorithms.

4.1 Signal Pre-processing

The recorded EEG signals are mostly, severely contaminated signals and are not actual brain signals. The contaminants are also known as artifacts. There are different kinds of artifacts such as power line noise, muscle contraction or electromyogram (EMG), heart activity or electrocardiogram (ECG), and eye movement or electrooculogram (EOG) [3]. These artifacts can be orders of magnitude

larger than the EEG signal. Therefore, the removal of artifacts is necessary to obtain the desired brain signals.

Many automated artifact removal methods have been proposed in the literature to remove artifacts from EEG recordings [18]. However, most of these methods either works well with additional EOG and EMG recordings or were designed to remove a single artifact. Hence, out of available artifact removal methods, we have used fully online and automated artifact removal tool for brain computer interfacing method (FORCe) [6] to remove all types of source generated artifacts. The clean data thus obtained (after the artifact removal phase) is further processed for baseline removal.

4.2 Channel Selection

Channel selection is done to choose the optimal subset of channels from the complete set of available channels. It is done to improve the model performance, provide faster processing, remove dimensionality curse, and to efficiently locate brain area that is responsible for neural activity.

In this work, we have used a very simple non-linear approach of channel selection which is called *Mutual Information (MI)*. It helps to evaluate non-linear dependencies between two or more random variables. Let X and Y be two random variables. Then, the *MI* between X and Y is the measure of amount of knowledge about Y which is provided by X and vice-versa. The *MI* between two random variables X and Y can be defined as:

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ I(Y; X) &= H(Y) - H(Y|X) \\ I(X; Y) &= H(X) + H(Y) - H(X; Y) \end{aligned} \quad (1)$$

where, $H(X)$ and $H(Y)$ are the entropies of random variables X and Y , and $H(X; Y)$ is their joint entropy. Their respective formulas are given here under.

$$\begin{aligned} H(X) &= - \int_X p_X(x) \log p_X(x) dx \\ H(Y) &= - \int_Y p_Y(y) \log p_Y(y) dy \\ H(X; Y) &= - \int_X \int_Y p_{X,Y}(x, y) \log p_{X,Y}(x, y) dx dy \end{aligned} \quad (2)$$

If *MI* between $H(X)$ and $H(Y)$ is zero, then X contains no information about random variable Y and vice-versa, which implies they are independent.

Based on *MI*, channels are either selected or rejected. We observed that AF3, F3, FC5, F7, F8, FC6, F4 and AF4 channels are confined to frontal lobe, which verifies the theory of cognition, according to which neuron activity related with cognitive workload is observed in the frontal lobe of human brain.

4.3 Feature Extraction

The feature extraction step involves extraction/selection of some distinctive components from the EEG signals. It is an extremely important step after signal

preprocessing and channel selection, as extraction of useful features is needed for classification of different levels of mental workload.

In this work, before extracting features, we divided the EEG data into epochs of length 3 s. Thus, for each epoch we obtained: $14 \text{ channels} \times 128 \text{ Hz} \times 3 \text{ s} = 14 \times 384 = 5376$ samples. Further, to classify the mental workload, we have calculated six different categories of features from the EEG signals, which are briefly discussed below:

- *Statistical features:* As EEG signal is a time-series signal and it can easily be characterized by the distribution of the amplitude and its statistical features. Therefore, for each epoch of an EEG signal, we calculated different statistical features and tabulated them in Table 1.

Table 1. Statistical features

| Features | Description |
|-----------|---|
| MEAN | Mean value |
| STD | Standard deviation |
| MAX_VALUE | Maximum positive amplitude |
| MIN_VALUE | Maximum negative amplitude |
| SKEWNESS | A measure of symmetry of the distribution |
| MEDIAN | The middle value of the set of ordered data |
| FD | Fractal dimension |
| AR | Auto regression |

- *Derivative features:* Derivative features are obtained by calculating the mean of first and second derivative of EEG signals and the maximum value of the first and second derivative of EEG signals. The extracted features are shown in Table 2.

Table 2. Derivative features

| Features | Description |
|---------------------------|--|
| 1 st DIFF_MEAN | Mean value of the first derivative of the signal |
| 1 st DIFF_MAX | Maximum value of the first derivative of the signal |
| 2 nd DIFF_MEAN | Mean value of the second derivative of the signal |
| 2 nd DIFF_MAX | Maximum value of the second derivative of the signal |

- *Interval or period features:* EEG signals can also be analyzed by measuring the distribution of the intervals between zero and other level crossings or between maxima and minima. The calculated interval features are listed in Table 3.

Table 3. Interval or period features

| Features | Description |
|--------------------|--|
| LINE_LENGTH | Line length |
| MEAN_VV_AMPL | Mean of vertex to vertex amplitudes |
| VAR_VV_AMPL | Variance of vertex to vertex amplitudes |
| MEAN_VV_TIME | Mean of vertex to vertex times |
| MEAN_VV_SLOPE | Mean of vertex to vertex slope |
| VAR_VV_SLOPE | Variance of vertex to vertex slope |
| ZERO_CROSSING | Number of zero crossings in the signal |
| MIN_MAX_NUMBER | Number of local minima and maxima |
| COEFF_OF_VARIATION | A statistical measure of the deviation of a variable from its mean, standard deviation divided by mean |
| AMPL_RANGE | The difference between maximum positive and maximum negative amplitude values |

- *Hjorth parameters:* Hjorth parameters gives an idea about the complexity of a time-series EEG signals. These values are very useful in EEG analysis and prove to be of great importance for its quantitative description. Refer Table 4 for the parameters.

Table 4. Hjorth parameters

| Features | Description |
|----------|--|
| HJORTH1 | Ability |
| HJORTH2 | Mobility $(\frac{\sigma_{x'}}{\sigma_x})$ |
| HJORTH3 | Complexity $(\frac{\frac{\sigma_{x''}}{\sigma_x'}}{\frac{\sigma_x}{\sigma_x'}})$ |

- *Frequency-domain features:* These features are one of the most important features for the analysis of EEG Signals. Based on the frequency content of the EEG signals, we extracted the features shown in Table 5 by applying the Fast Fourier Transform (FFT) to various EEG wave bands. Further, we also calculated other important ratios of FFT from various bands.
- *Wavelet features:* The wavelet transform (WT) is capable of distinguishing very small and delicate differences between time-series signals even from short signal epochs. It can easily identify highly irregular and non-stationary signals. Further, WT based methods can localize the signal components in time-frequency space in a better way than FFT analysis. Therefore, we evaluated the features listed in Table 6 using WT.

Table 5. Frequency-domain features

| Features | Description |
|-----------------|---|
| FFT_DELTA | 0.1–4 Hz |
| FFT_THETA | 4–8 Hz |
| FFT_ALPHA | 8–13 Hz |
| FFT_BETA | 13–30 Hz |
| FFT_GAMMA | 30–40 Hz |
| FFT_WHOLE | .1–40 Hz |
| FFT_DT_RATIO | DELTA/THETA |
| FFT_DA_RATIO | DELTA/ALPHA |
| FFT_TA_RATIO | THETA/ALPHA |
| FFT_DTA_RATIO | (DELTA+THETA)/ALPHA |
| FFT_SEF | Spectral edge frequency |
| FFT_SP_ROLL_OFF | Below which 85% of the total spectral power resides |

Table 6. Wavelet features

| Features | Description |
|--------------------------------|--|
| MIN_WAV_VALUE | Minimum value |
| MAX_WAV_VALUE | Maximum value |
| MEAN_WAV_VALUE | Mean value |
| MEDIAN_WAV_VALUE | Median value |
| STD_WAV_VALUE | Standard deviation |
| SKEWNESS_WAV_VALUE | Skewness |
| KURTOSIS_WAV_VALUE | Kurtosis |
| WAV_BAND | Relative energy |
| ENTROPY_SPECTRAL_WAV | The spectral entropy |
| 1 st _DIFF_WAV_MEAN | Mean value of the 1 st derivative |
| 1 st _DIFF_WAV_MAX | Maximum value of the 1 st derivative |
| 2 nd _DIFF_WAV_MEAN | Mean value of the 2 nd derivative |
| 2 nd _DIFF_WAV_MAX | Maximum value of the 2 nd derivative |
| ENERGY_PERCENT_WAV | Percentage of the total energy of a detail/approximation |
| WAV_ZERO_CROSSING | Zero crossing |
| WAV_COEFF_OF_VARIATION | Coefficient of variation |
| WAV_TOTAL_ENERGY | Total energy |

4.4 Feature Normalization and Optimization

The extracted features are normalized to bring them within a common range. This helps in feature optimization and reduces the inter-subject variability. Here, we have mean-normalized the extracted features using Eq. 3.

$$x_{new} = \frac{x - \mu}{\sigma} \tag{3}$$

where, μ and σ denote *mean* and *standard deviation*, respectively.

Feature optimization also helps in minimizing the curse of dimensionality and enhanced generalization by reducing over-fitting. In feature optimization/selection, we identify data that are relevant to the selected parameters and assign them *maximum relevance*. We select those features which are strongly correlated to the classification and call this task as *maximum-relevance selection*. Besides, features which are mutually separated but have high degree of correlation to the classification are also selected and this task is referred to as *minimum-redundant selection*. These parameters are sometimes redundant and can be easily suppressed using maximum Relevance Minimum Redundancy (mRMR) algorithm [15]. Therefore, we have applied the mRMR algorithm to the extracted feature set to obtain the most optimized set of features. The features obtained after applying the optimization algorithm are tabulated in Table 7.

Table 7. Optimized features

| Feature | Description |
|----------------------------|---|
| FD | Fractal dimension |
| AR | Auto regression |
| 1 st _DIFF_MEAN | Mean value of the first derivative of the signal |
| 1 st _DIFF_MAX | Maximum value of the first derivative of the signal |
| 2 nd _DIFF_MEAN | Mean value of the second derivative of the signal |
| 2 nd _DIFF_MAX | Maximum value of the second derivative of the signal |
| HJORTH1 | Ability |
| HJORTH2 | Mobility ($\frac{\sigma'}{\sigma_x}$) |
| HJORTH3 | Complexity ($\frac{\frac{\sigma''}{\sigma_x}}{\frac{\sigma'}{\sigma_x}}$) |
| FFT_DT_RATIO | $\frac{DELTA}{THETA}$ |
| FFT_DA_RATIO | $\frac{DELTA}{ALPHA}$ |
| FFT_TA_RATIO | $\frac{THETA}{ALPHA}$ |
| FFT_DTA_RATIO | $\frac{DELTA+THETA}{ALPHA}$ |
| WAV_COEFF_OF_VARIATION | Coefficient of variation |
| WAV_TOTAL_ENERGY | Total energy |

5 Results and Discussion

In this section, we present the spectrogram plots for the five levels of workload data. Next, we show the classification accuracy results for pre and post channel selection and feature extraction, respectively. At last, we present confusion matrix for pre and post channel selection and feature extraction, respectively.

For easy identification, we have labelled our cognitive workload level as C_i , where $i = \{1,2,3,4,5\}$, wherein C_1 denotes *idle* task, C_2 denotes 1-back task, C_3 denotes 2-back task, C_4 denotes dual 1-back task and C_5 denotes dual 2-back task.

Various combinations of two-class, three-class, four-class and five-class classification for the above-mentioned categories have been summarized in the form of table and bar-chart. The obtained results are described next.

5.1 Spectrogram Plot

We have plotted spectrograms for all (five) levels of cognitive tasks for subject *M05*. From these plots (see Figs. 4, 5, 6, 7 and 8) we find/visualize the dominant EEG bands in a cognitive task. It can be clearly seen that *theta* and *alpha* wave activities are the most dominant in these spectrograms and possess most of the band power. Moreover, from the spectrograms for dual 2-back task one can notice that there is an increase in the beta band activity due to an increase in cognitive workload.

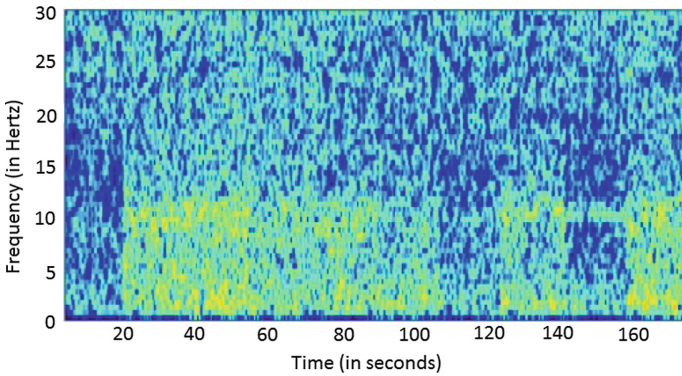


Fig. 4. Spectrogram plot for idle task

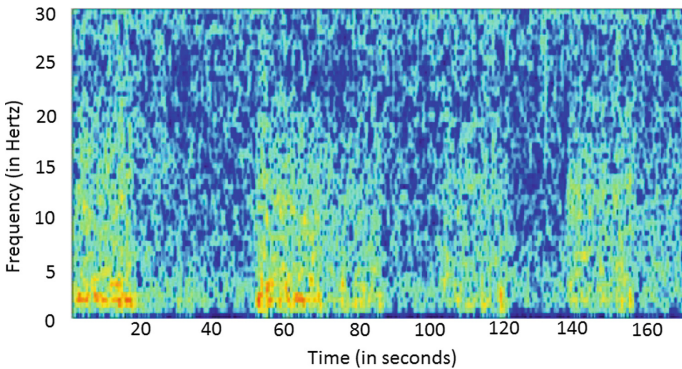


Fig. 5. Spectrogram plot for 1-back task

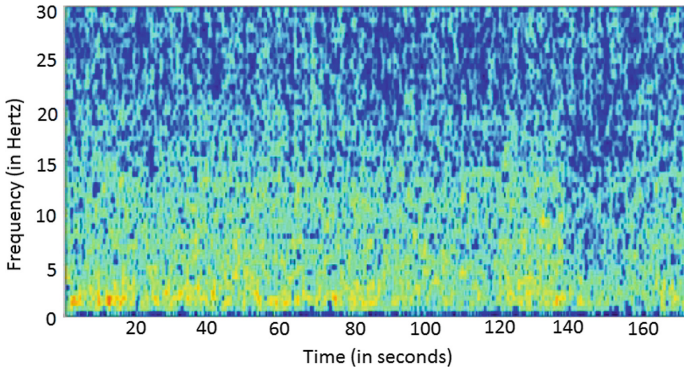


Fig. 6. Spectrogram plot for 2-back task

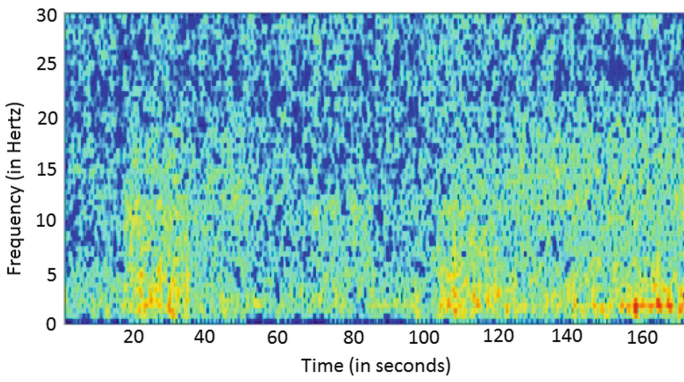


Fig. 7. Spectrogram plot for dual 1-back task

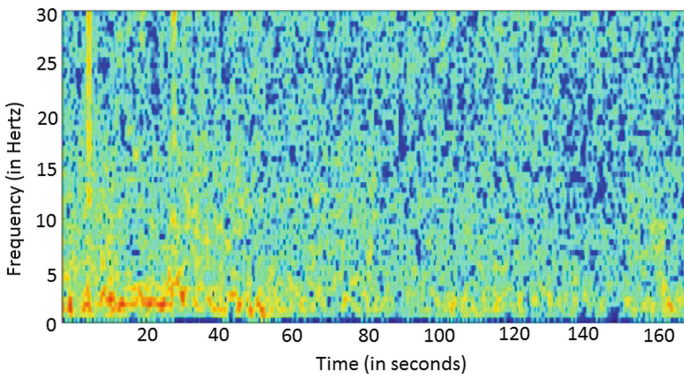


Fig. 8. Spectrogram plot for dual 2-back task

5.2 Classification Accuracy Using the Classifiers

Classification of the mental workload data into different levels has been done with the aid of seven supervised machine learning algorithms already named in Sect. 1. Each classifier model has been trained by dividing the complete dataset into a training set comprising of 80% data values and a test set comprising of remaining 20% data values. We have used *scikit-learn* open library for executing our machine learning algorithms. Further, for visualizing the effect of channel selection and feature optimization, we have carried out our classification in two different categories which are described next. In addition, we have also summarized various combinations of two-class, three-class, four-class and five-class classifications and summarized them in the form of Tables 8 and 9.

Table 8. Classification accuracy in (%) without channel selection and feature optimization

| Class | Classifier | | | | | | |
|--|------------|---------------|-----------|--------------|----------|----------|---------|
| Two-class | k-NN | Random forest | SVM (RBF) | SVM (Linear) | MLP | D-Tree | LDA |
| C ₁ -C ₂ | 86.55 | 97.47 | 89.00 | 94.00 | 94.95 | 87.39 | 92.01 |
| C ₁ -C ₃ | 90.80 | 97.99 | 94.00 | 89.00 | 96.39 | 90.80 | 90.40 |
| C ₁ -C ₄ | 92.40 | 96.39 | 94.00 | 92.00 | 94.39 | 90.00 | 91.20 |
| C ₁ -C ₅ | 89.55 | 95.58 | 91.00 | 86.00 | 93.57 | 91.16 | 85.14 |
| C ₂ -C ₃ | 81.09 | 89.91 | 83.00 | 86.00 | 89.70 | 77.31 | 81.51 |
| C ₂ -C ₄ | 92.85 | 95.37 | 88.00 | 89.00 | 93.69 | 86.55 | 92.43 |
| C ₂ -C ₅ | 84.81 | 92.40 | 84.00 | 91.00 | 92.82 | 81.85 | 88.60 |
| C ₃ -C ₄ | 77.60 | 84.39 | 74.00 | 72.00 | 82.39 | 73.99 | 78.40 |
| C ₃ -C ₅ | 81.92 | 86.74 | 75.00 | 75.00 | 82.32 | 73.49 | 79.91 |
| C ₄ -C ₅ | 83.13 | 86.34 | 80.00 | 84.00 | 88.35 | 77.91 | 83.93 |
| Average | 86.07 | 92.258 | 85.2 | 85.8 | 90.857 | 83.045 | 86.353 |
| Three-class | k-NN | Random forest | SVM (RBF) | SVM (Linear) | MLP | D-Tree | LDA |
| C ₁ -C ₂ -C ₃ | 77.41 | 91.46 | 73.00 | 82.00 | 84.29 | 79.06 | 82.09 |
| C ₁ -C ₃ -C ₅ | 76.47 | 87.96 | 73.00 | 75.00 | 82.08 | 68.71 | 71.92 |
| C ₂ -C ₃ -C ₄ | 66.94 | 79.33 | 62.00 | 70.00 | 75.75 | 67.49 | 70.79 |
| C ₃ -C ₄ -C ₅ | 66.84 | 82.62 | 65.00 | 66.00 | 71.65 | 66.04 | 67.11 |
| Average | 71.915 | 85.3425 | 68.25 | 73.25 | 78.4425 | 70.325 | 72.9775 |
| Four-class | k-NN | Random forest | SVM (RBF) | SVM (Linear) | MLP | D-Tree | LDA |
| C ₁ -C ₂ -C ₃ -C ₄ | 65.98 | 81.55 | 64.00 | 68.00 | 76.02 | 67.82 | 71.31 |
| C ₂ -C ₃ -C ₄ -C ₅ | 67.14 | 79.87 | 60.00 | 63.00 | 72.07 | 60.36 | 62.42 |
| C ₁ -C ₂ -C ₄ -C ₅ | 71.66 | 86.44 | 70.00 | 75.00 | 82.54 | 68.58 | 76.79 |
| Average | 68.26 | 82.62 | 64.6667 | 68.6667 | 76.87667 | 65.58667 | 70.1733 |
| Five-class | k-NN | Random forest | SVM (RBF) | SVM (Linear) | MLP | D-Tree | LDA |
| C ₁ -C ₂ -C ₃ -C ₄ -C ₅ | 61.43 | 80.22 | 57.00 | 62.00 | 68.13 | 58.33 | 63.39 |

Table 9. Classification accuracy in (%) with channel selection and feature optimization

| Class | Classifier | | | | | | |
|--|------------|---------------|-----------|--------------|----------|---------|----------|
| | k-NN | Random forest | SVM (RBF) | SVM (Linear) | MLP | D-Tree | LDA |
| Two-class | k-NN | Random forest | SVM (RBF) | SVM (Linear) | MLP | D-Tree | LDA |
| C ₁ -C ₂ | 95.19 | 99.19 | 95.00 | 88.00 | 95.99 | 92.00 | 95.19 |
| C ₁ -C ₃ | 93.60 | 95.19 | 96.00 | 95.00 | 94.39 | 91.20 | 96.79 |
| C ₁ -C ₄ | 95.19 | 93.60 | 97.00 | 97.00 | 99.19 | 88.00 | 97.59 |
| C ₁ -C ₅ | 94.35 | 95.96 | 94.00 | 92.00 | 96.77 | 87.90 | 91.12 |
| C ₂ -C ₃ | 83.19 | 92.80 | 82.00 | 94.00 | 86.39 | 85.59 | 91.20 |
| C ₂ -C ₄ | 95.19 | 96.79 | 90.00 | 92.00 | 92.00 | 91.20 | 90.40 |
| C ₂ -C ₅ | 86.29 | 94.35 | 88.00 | 95.00 | 93.54 | 84.67 | 91.12 |
| C ₃ -C ₄ | 81.59 | 83.99 | 75.00 | 76.00 | 80.00 | 74.39 | 80.80 |
| C ₃ -C ₅ | 79.03 | 88.70 | 78.00 | 76.00 | 76.61 | 75.00 | 76.61 |
| C ₄ -C ₅ | 80.64 | 87.90 | 82.00 | 87.00 | 83.06 | 81.45 | 83.87 |
| Average | 88.426 | 92.928 | 87.7 | 89.2 | 89.794 | 85.14 | 89.469 |
| Three-class | k-NN | Random forest | SVM (RBF) | SVM (Linear) | MLP | D-Tree | LDA |
| C ₁ -C ₂ -C ₃ | 83.51 | 91.48 | 86.00 | 89.00 | 88.82 | 82.97 | 86.70 |
| C ₁ -C ₃ -C ₅ | 81.28 | 86.63 | 83.00 | 83.00 | 87.16 | 75.40 | 86.09 |
| C ₂ -C ₃ -C ₄ | 75.00 | 91.48 | 71.00 | 78.00 | 80.85 | 82.44 | 77.65 |
| C ₃ -C ₄ -C ₅ | 66.31 | 79.67 | 66.00 | 69.00 | 76.47 | 65.77 | 74.33 |
| Average | 76.525 | 87.315 | 76.5 | 79.75 | 83.325 | 76.645 | 81.1925 |
| Four-class | k-NN | Random forest | SVM (RBF) | SVM (Linear) | MLP | D-Tree | LDA |
| C ₁ -C ₂ -C ₃ -C ₄ | 71.59 | 87.60 | 73.00 | 77.00 | 85.19 | 77.20 | 79.20 |
| C ₂ -C ₃ -C ₄ -C ₅ | 67.87 | 82.73 | 64.00 | 73.00 | 73.09 | 69.07 | 78.71 |
| C ₁ -C ₂ -C ₄ -C ₅ | 74.29 | 90.76 | 74.00 | 84.00 | 89.15 | 73.89 | 82.32 |
| Average | 71.25 | 87.03 | 70.3333 | 78 | 82.47667 | 73.3867 | 80.07667 |
| Five-class | k-NN | Random forest | SVM (RBF) | SVM (Linear) | MLP | D-Tree | LDA |
| C ₁ -C ₂ -C ₃ -C ₄ -C ₅ | 64.42 | 84.61 | 63.00 | 73.00 | 79.80 | 69.55 | 78.20 |

Classification without Channel Selection and Feature Optimization:

From the obtained results (refer Table 8) it can be observed that the Random Forest algorithm gives the best classification accuracy for all combinations of classes. It can be noted that, this classifier presents highest accuracy of 97.22% for two-class classification followed by percentage accuracy of 91.46, 86.44 and 80.22 for the combination of three, four and five classes, respectively.

Classification with Channel Selection and Feature Optimization:

After channel selection and feature optimization, it has been observed that the average classification accuracy increases for all the classifiers involved (refer Table 9). Further, it has been observed that the Random Forest classifier outperforms all other classifiers. Highest accuracy obtained with Random Forest is 99.19% in two-class classification followed by percentage accuracy of 91.48%, 90.76% and 84.61% for three, four and five classes, respectively.

5.3 Confusion Matrix

For efficiently depicting the accuracy of classification, we have shown the obtained results with the help of confusion matrix. Each column of the matrix represents the instances in a predicted class while each row represents the instances in a true class (or vice-versa). The diagonal elements represent the number of points for which the predicted class is same as the true class, while off-diagonal elements are those which are misclassified by the classifier. Higher diagonal values of the confusion matrix indicates better predictions. We present confusion matrix in two categories which are discussed next.

Confusion Matrix Before Channel Selection and Feature Optimization: Confusion matrix before channel selection and feature optimization techniques for two-class, three-class, four-class and five-class classification is shown in Fig. 9.

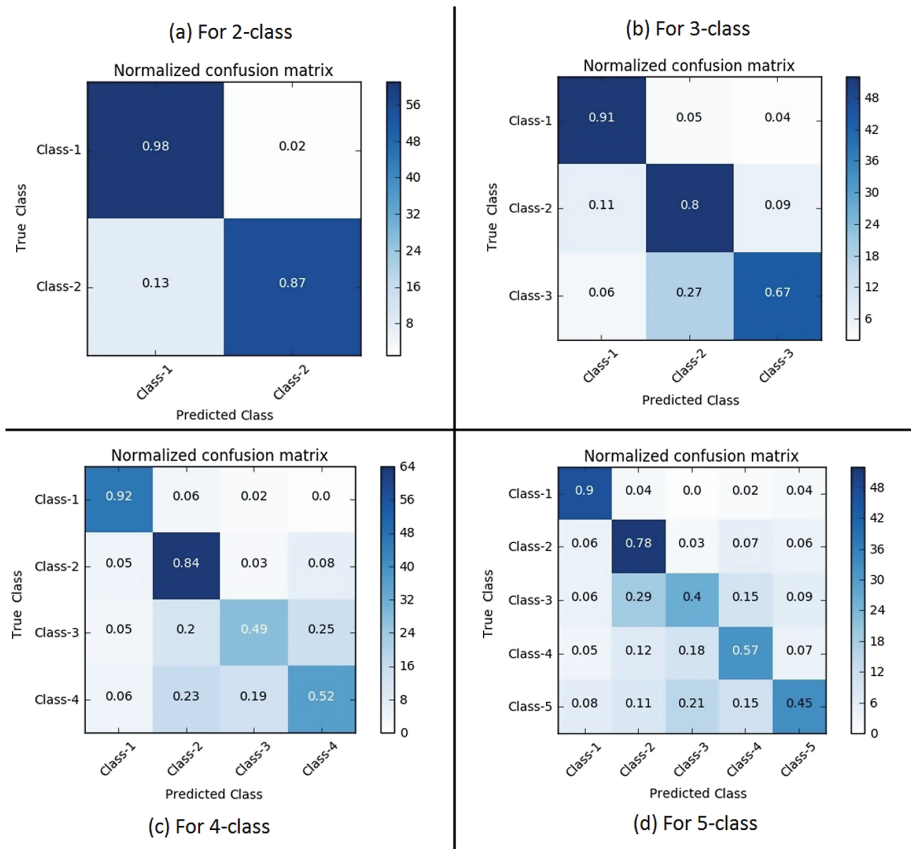


Fig. 9. Confusion matrix before channel selection and feature optimization:

Confusion Matrix After Channel Selection and Feature Optimization:

Confusion matrix after channel selection and feature optimization techniques for two-class, three-class, four-class and five-class classification is shown in Fig. 10.

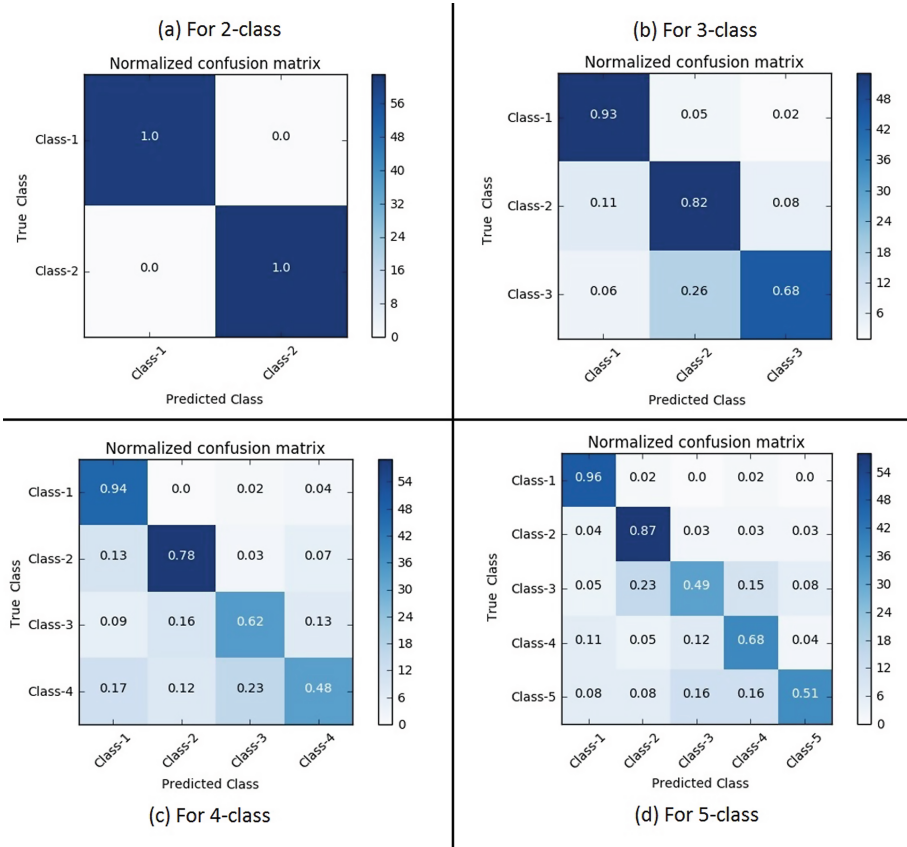


Fig. 10. Confusion matrix after channel selection and feature optimization

On comparing the matrices for the two cases, we can observe that there is a substantial improvement in classification accuracy with the usage of channel selection and feature optimization for all class labels. For instance, we can note that, in class-1 to class-1 matching from both cases, the accuracy increases from 98% to 100% in two class classification, 91% to 93% in three class classification, from 92% to 94% in four class classification and from 90% to 96% in five class classification.

6 Conclusion

In this paper, first, we explored the feasibility of wireless data acquisition devices in mental workload assessment with the help of n -back task. From this it is evident that these devices have enormous potential which may be exploited in everyday environment and can be of utmost importance to handle critical situations such as monitoring pilots of flights, nuclear operations, driving tasks, etc. Second, we modeled and evaluated MWL induced during human-computer interaction with the help of features extracted from EEG signals. Third, we investigated the potential of machine learning to classify MWL in different levels. To accomplish this, we have used different categories of supervised machine learning algorithms that can learn from the data (about its pattern) and give predictions. Fourth, we studied the effect of channel selection and feature optimization on classification performance. From the obtained results, it can be easily observed that the Random Forest algorithm results in best accuracy in comparison to all the other compared algorithms. Further, we also studied the performance accuracy obtained due to inter-class classification. We hope that this study would be helpful in future to explore and devise new methods for studying and understanding cognitive workload.

References

1. Anderson, E.W., Potter, K.C., Matzen, L.E., Shepherd, J.F., Preston, G.A., Silva, C.T.: A user study of visualization effectiveness using EEG and cognitive load. *Comput. Graph. Forum* **30**(3), 791–800 (2011)
2. Ayaz, H., Onaral, B., Izzetoglu, K., Shewokis, P.A., McKendrick, R., Parasuraman, R.: Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: empirical examples and a technological development. *Front. Human Neurosci.* **7**, 871 (2013)
3. Benbadis, S.R.: EEG artifacts. <http://emedicine.medscape.com/article/1140247-overview#a3>. Accessed 14 Feb 2018
4. Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L.: EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Sp. Environ. Med.* **78**(5), B231–B244 (2007)
5. Chaouachi, M., Jraidi, I., Frasson, C.: Modeling mental workload using EEG features for intelligent systems. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) *UMAP 2011*. LNCS, vol. 6787, pp. 50–61. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22362-4_5
6. Daly, I., Scherer, R., Billinger, M., Müller-Putz, G.: Force: fully online and automated artifact removal for brain-computer interfacing. *IEEE Trans. Neural Syst. Rehabil. Eng.* **23**(5), 725–736 (2015)
7. Heine, T., Lenis, G., Reichensperger, P., Beran, T., Doessel, O., Deml, B.: Electrocardiographic features for the measurement of drivers' mental workload. *Appl. Ergon.* **61**, 31–43 (2017)

8. Hirshfield, L.M., Chauncey, K., Gulotta, R., Girouard, A., Solovey, E.T., Jacob, R.J.K., Sassaroli, A., Fantini, S.: Combining electroencephalograph and functional near infrared spectroscopy to explore users' mental workload. In: Schmorow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) FAC 2009. LNCS (LNAI), vol. 5638, pp. 239–247. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02812-0_28
9. Holm, A., Lukander, K., Korpela, J., Sallinen, M., Müller, K.M.I.: Estimating brain load from the EEG. *Sci. World J.* **9**, 639–651 (2009)
10. Hoskinson, P.: Brain workshop - a dual n-back game. <http://brainworkshop.sourceforge.net/>. Accessed 14 Feb 2018
11. Ke, Y., Qi, H., He, F., Liu, S., Zhao, X., Zhou, P., Zhang, L., Ming, D.: An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. *Front. Hum. Neurosci.* **8**, 703 (2014)
12. Mahmoud, R., Shanableh, T., Bodala, I.P., Thakor, N., Al-Nashash, H.: Novel classification system for classifying cognitive workload levels under vague visual stimulation. *IEEE Sens. J.* **17**, 7019–7028 (2017)
13. Moustafa, K., Luz, S., Longo, L.: Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 30–50. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_3
14. Mühl, C., Jeunet, C., Lotte, F.: EEG-based workload estimation across affective contexts. *Front. Neurosci.* **8**, 114 (2014)
15. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1226–1238 (2005)
16. Roy, R.N., Bonnet, S., Charbonnier, S., Campagne, A.: Efficient workload classification based on ignored auditory probes: a proof of concept. *Front. Hum. Neurosci.* **10**, 519 (2016)
17. Samima, S., Sarma, M., Samanta, D.: Correlation of P300 ERPS with visual stimuli and its application to vigilance detection. *IEEE*, July 2017
18. Urigüen, J.A., Garcia-Zapirain, B.: EEG artifact removal-state-of-the-art and guidelines. *J. Neural Eng.* **12**(3), 031001 (2015)
19. Wang, S., Gwizdka, J., Chaovaitwongse, W.A.: Using wireless EEG signals to assess memory workload in the *n*-back task. *IEEE Trans. Hum.-Mach. Syst.* **46**(3), 424–435 (2016)
20. Zarjam, P., Epps, J., Chen, F., Lovell, N.H.: Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Comput. Biol. Med.* **43**(12), 2186–2195 (2013)



Psycho-Physiological Evaluation of the Pilot: A Study Conducted with Pilots of the French Air Force

Vincent Ferrari¹, Jean-François Gagnon^{2(✉)}, Cyril Camachon¹, and Maëlle Kopf³

¹ Air Force Research Center, Salon-de-Provence, France
vincent.ferrari@ecole-air.fr

² Thales Canada Inc. Research and Technology, Québec, Canada
jean-francois.gagnon@ca.thalesgroup.com

³ Thales AVS, Osny, France

Abstract. The present paper concerns the individualization of the training of aircraft pilots. Specifically, it presents the data collection, and modeling efforts carried out to assess trainees' transition from a controlled, effortful piloting experience (i.e., System 2), to an automatic, effortless process (i.e., System 1). It is argued that cardiovascular activity can be associated with deployment of effort, and therefore be used for assessing the transition across systems. Heart rate, respiration rate, and heart rate variability were sampled on 11 pilots (5 students "novice" and 6 instructors "experts"), performing 6 one-hour flights (5 flights in tandem: one student and one instructor, the 6th flight with an instructor flying alone). These data were used for the development of a prediction model computing the probability of a pilot being an expert or a novice. After a "leave-one-tandem-out" validation, the accuracy of the model was 86.86%. The results are discussed in terms of effortful processes and skill acquisition. Further work will consist in implementing contextual parameters in the model in order to improve the prediction. Such a model could be used by instructors and trainees as a supporting tool for tracking progress of the training at the individual level.

Keywords: Psycho-physiology · Heart rate · Heart rate variability · Ease of flight Training · Quantification of learning

1 Introduction

In a complex and dynamic environment such as piloting a fighter plane, a pilot must constantly be prepared to react to unexpected situations, and engage additional cognitive resources for carrying out his mission. Hence, an important part of military pilot training is to appreciate the unpredictable nature of the mission (Fornette et al. 2015). This educational approach is all the more demanding since it begins early in the formation of the trainee and should allow him to perform well in spite of the uncertainty of the operating environment.

For the instructor, identifying the specific moment to go from an expected to an unexpected situation is crucial, as the change must only be operated once the trainee has acquired the fundamentals of the flight, that is to say a set of knowledge devoted to maintaining the aircraft in a safe area (i.e. maintain altitude, heading, etc.). If unexpected

changes are brought too early in the training, the trainee will not be able to fully integrate the basis, and will not reach the optimal and target state which is referred to as “ease in flight”. Indeed, excessive demand on resources imposed by the attended task(s) typically results in performance degradation (Nourbakhsh et al. 2013; Stanton et al. 2005). On the contrary, a delayed addition of unexpected situations will have no benefit and inefficiently extend the training time. Currently, instructors exclusively rely on their experience and subjective observations to detect this key moment.

From a cognitive point of view, ease of flight could be associated with automatic processes (i.e., System 1) as opposed to controlled processes (i.e., System 2). Over the last decades, this multiple system theory of decision making has been widely studied and has accumulated a large body of evidence (see Sanfey and Chang 2008 for a brief review). System 1 has been described as fast, effortless, and unconscious whereas system 2 has been depicted as slow, effortful, and conscious.

Skill acquisition can be viewed as a shift from system 2 to system 1. Kahneman (2003) has even linked System 1 to “intuition”, frequently associated with how experts make decisions (e.g., Dreyfus 2014 [in Zsombok and Klein]).

Automated processes require very little cognitive resources, as opposed to controlled processes. An expert pilot has automated the majority of recurrent piloting tasks and procedures (e.g., take-off), which, for him, do not require an important engagement of cognitive resources. In comparison, a novice who has not fully automated the procedures will have to spend more energy to reach a similar performance. In a systemic view of the phenomenon, this difference in terms of energetic cost is expected to have physiological corollaries, in particular cardiorespiratory, which can be used as indicators of energetic spending. Indeed, several physiological corollaries of cognitive efforts have been identified over the last decades, including in piloting tasks (Roscoe 1992).

For instance, increase in heart rate (HR) has been associated with effort, cognitive (e.g., Kennedy and Scholey 2000) and physical. Notably, it was used by Dahlstrom and Nahlinder (2009) to estimate mental workload for pilots in simulators and in-flight. It has great potential for in-flight mental workload estimation because it is easily obtained, and less subject to noise than other typically used measures, like electro-encephalogram. HR variability (HRV) refers to the regularity of consecutive R-R intervals of the QRS complex as measured by an electro-cardiogram (ECG). Although not as intuitive as HR, HRV is one of the most frequently used metric associated with mental effort, both in fundamental and applied research. For instance, HRV was associated with mental overload in a simulated piloting task (Durantin et al. 2014), and with several fundamental neuro-cognitive tasks (Gagnon et al. 2016). Finally, respiration rate (RR) has been linked with energetic spending, has been considered a measure of task demands (Overbeek et al. 2014) and was also associated with negative valance and arousal (Masa et al. 2003).

In the context of air force pilot training, we hypothesize that during identical flights, the trainees will have to deploy a greater amount of mental effort than instructors for reaching similar performances. Therefore, trainees should exhibit a specific pattern of physiological parameters: HR and RR should be higher, and HRV should be lower than for instructors. Based on this premise, we hypothesize that it is possible to predict the role of the pilot (trainee or instructor) using physiological measures.

Moreover, the use of physiological measures could allow the identification of “expected pattern” among experienced pilots, which would be used as references when considering the same metrics among trainees in identical situations. The variation of the difference between the expected pattern (expert) and the observation (trainee) could be interpreted as a consequence of the levels of cognitive automation of the processes in the given situation for a trainee. Hence, this paper considers the possibility of quantifying learning by comparing his metrics to the reference measured on his instructor.

1.1 Objectives

The main goal of the present paper is to open the way towards an objective measure of “ease in flight”, which would assist instructors and their students during training. Such an objective measure would be a key element in the process of individualizing the training of pilots. As learning skills have a great variability between trainees, objectively quantifying to which degree a student easily performs a task could allow a great improvement in the training. Specifically, this paper is organized around two objectives, described below.

Objective 1

The first objective is to assess the impact of roles and flight phases on physiological measures. Specifically, three variation of the main hypothesis are formulated:

- H1. Mean physiological values will differ across roles
 - H1a. Mean heart rate will be higher for trainees when compared with instructors
 - H1b. Mean heart rate variability will be lower for trainees when compared with instructors
 - H1c. Breathing rate should be higher for trainees when compared with instructors

Objective 2

The second objective is to develop a model for predicting the level of expertise based on physiological measures. The physiological predictors will be comprised of statistically significant predictors that varied across roles. The model will be applied to the physiological measures and predicted expertise will be assessed. This model assumes that instructors have greater expertise than trainees.

The goal is to evaluate if such a model could help dynamically (1) quantify the progression of training and (2) identify periods of time where the instructor might not be fully in control of the flight.

2 Method

Eleven pilot participants were equipped with a Zephyr Bio Harness 3.0 chest strap measuring the electrical activity of the heart (ECG), RR, and accelerations on 3-axis. They were also equipped with an Android mobile phone on which the Sensor Hub (Gagnon et al. 2016) application was installed. The application integrates all generated data, processes HR, HRV (frequency and temporal domain), accelerations (3-axis), respiration rate, and global positioning system coordinates.

Participants were organized in tandems consisting of an instructor (assumed expert) and a trainee (assumed novice). The data were collected on five comparable aerobic flights with trainees of approximately the same skill level. One of the flights was performed by an instructor flying alone. Each flight was broken down into five phases: pre-flight (briefing), take-off, flight, landing, and post-flight (debriefing).

During the flight, instructors performed specific maneuvers that the trainees had to perform immediately after, therefore transferring the control of the plane from one to another. Instructors were responsible for take-off and landing.

3 Results

Results are described in two sub-sections, aligned with the objectives. First the statistical significance tests are reported to evaluate the impact of the key factors (role and phases) on individual physiological measures. Second, a classifier of expertise is developed and described.

3.1 Factors Influencing Physiological Parameters

Three mixed ANOVAs were carried out to test the effect of the role (Trainee vs Instructor), phase (repeated 5 levels), and their interaction on (1) mean HR in bpm, (2) mean HRV in ms, and (3) mean RR in bpm.

Hypothesis H1a

Results show that both role $F(1,6) = 9.27, p < .05$ and phase $F(4,30) = 14.51, p < .001$ had a statistically significant impact on mean HR in bpm. Interaction of role and phase was not statistically significant $F(4,30) = 1.81, N.S$. In line with hypothesis H1a, mean HR in bpm is statistically higher in the trainee condition (mean = 113.56, sd = 22.73) when compared with the instructor condition (mean = 74.82, sd = 11.21). Results are presented in Fig. 1.

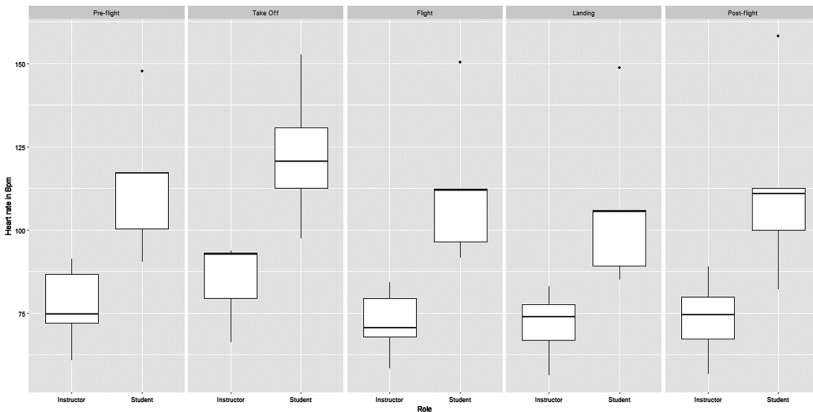


Fig. 1. Mean HR in bpm by role and phase.

Hypothesis H1b

Results show that both role $F(1,6) = 11.30, p < .05$ and phase $F(4,30) = 3.50, p < .05$ had a statistically significant impact on mean HRV in ms. Interaction of role and phase was not statistically significant $F(4,30) = 1.77, N.S.$ In line with hypothesis H1b, mean HRV in ms is statistically lower in the student condition (mean = 36.58, sd = 16.93) when compared with the instructor condition (mean = 65.66, sd = 17.80). Results are presented in Fig. 2.

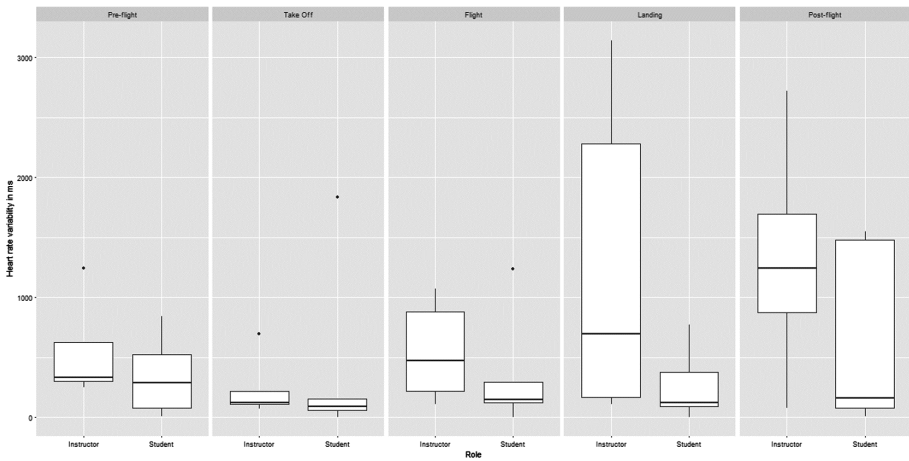


Fig. 2. Mean HRV by role and phase.

Hypothesis H1c

Results show that phase $F(4,30) = 5.48, p < .001$ had a statistically significant impact on mean RR in bpm. Role did not have a significant impact $F(1,6) = 1.11, N.S.$ However interaction of role and phase was statistically significant $F(4,30) = 4.39, p < .01$. Unsupportive of hypothesis H1c, mean RR in bpm is not statistically higher in the student condition (mean = 19.80, sd = 2.21) when compared with the instructor condition (mean = 18.23, sd = 3.26), but there is a significant interaction of the two factors $F(4,30) = 4.66, p < .01$ on respiration rate. Results are presented in Fig. 3.

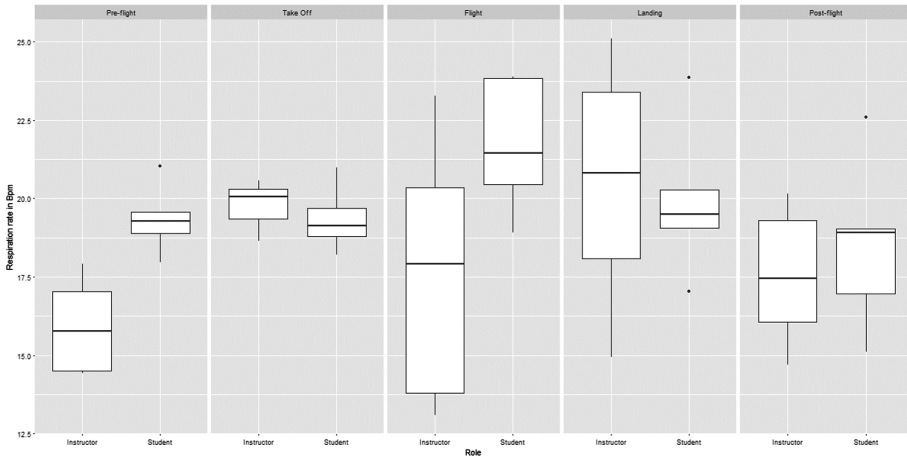


Fig. 3. Mean respiration rate in bpm by role and phase.

3.2 Modeling Effort Linked with Expertise

In addition to statistical significance tests, an integrated model was developed to predict the role of the participant based on HR, HRV, and RR as predictors. In an attempt to remain parsimonious and explainable, the generalized linear model (GLM) was employed. However, rather than using phases as temporal separations, equal non-overlapping bins of 10 min were created. For each of these bins, mean HR in bpm, mean HRV in ms and mean RR in bpm were calculated. The model was developed using this data. The reason for the creation of bins is that the flight phases are highly variable in terms of length and would therefore induce a bias in the statistical representativeness of metrics within the model. For instance, a very short phase of two minutes would have the same statistical weight than a phase lasting 45 min.

The model was validated for generalization using a “leave-one-tandem-out” procedure. The final model used for predictions was retrained on all the data.

Results show that the model achieved an accuracy of 86.86% (95% confidence interval = 80.03–92.02), $\chi = .74$. The predictors (and associated betas β) are represented in order of relative influence in Table 1.

Table 1. Model predictors and associated β .

| Predictor | β |
|------------------------------|---------|
| Heart rate in bpm | -2.6211 |
| Heart rate variability in ms | 1.1584 |
| Breathing rate in bpm | 0.4858 |
| Intercept | -0.1214 |

The model was then applied to each individual data to see how the predictions unfold in time during a flight. The numeric prediction represents the probability that the

observed physiological pattern (composed of HR, HRV, and RR) is the one of an instructor. Hence, when the probability exceeds 50%, the point is classified as “instructor”, and conversely when below 50%. By showing the predicted probability, we can track changes in the progression of each individual. We plotted the predictions of two tandems that were deemed interesting for discussion. Tandem 1 model predictions were plotted in Fig. 4, and tandem 5 in Fig. 7. Alongside the predictions of the model, we plotted the most influencing factor of the model (i.e., heart rate in bpm), and altitude in meters to provide some context.

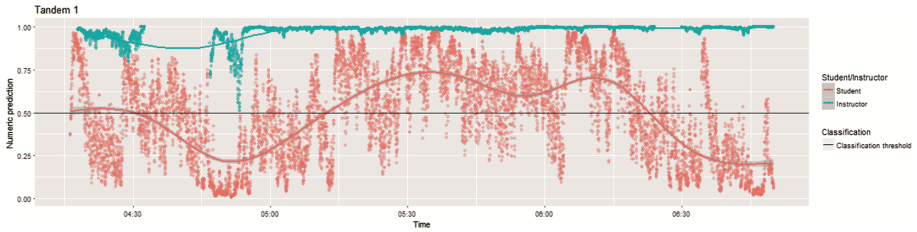


Fig. 4. Tandem 1 - Probability of being an instructor according to the model, by role for the whole flight. The classification threshold corresponds to the point where the most probable classification changes from one role to another. The points over the horizontal line (Classification threshold, 0.5) represent the data which were classified as being those of an instructor.

Tandem 1 (Figs. 4, 5 and 6) shows that the instructor was classified as an instructor all the time. Interestingly, results show that the student was above the 50% threshold (so classified as an instructor) for a long period of the flight, but still had punctual states corresponding to the typical state of a “trainee”.

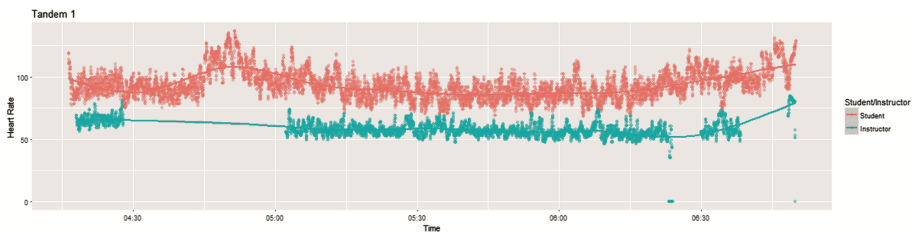


Fig. 5. Tandem 1 - HR in bpm sampling values through the flight. The predictions made by the model are largely based on this metric.

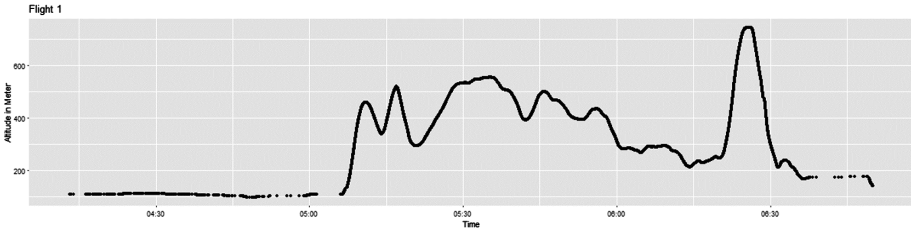


Fig. 6. Tandem 1 - Altitude in meters of the aircraft.

Tandem 5 (Figs. 7, 8 and 9) resulted in a much different pattern than the previous tandem. First, it is observed that the instructor is not classified with as much confidence as instructor from Tandem 1. Punctually, the probability of being an instructor even falls below the 50% threshold. On the other hand, the data shows a progression of the trainee from trainee to instructor as the flight progresses.

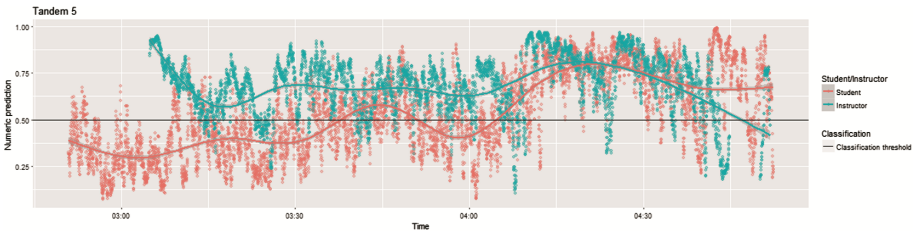


Fig. 7. Tandem 5 - Probability of being an instructor according to the model, by role for the whole flight. The classification threshold corresponds to the point where the most probable classification changes from one role to another. The points over the horizontal line (Classification threshold, 0.5) represent the data which were classified as being those of an instructor.

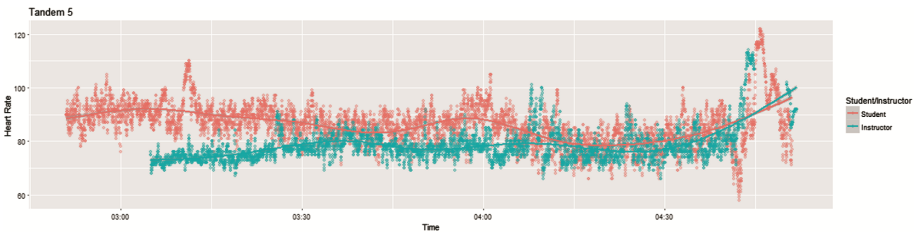


Fig. 8. Tandem 5 – HR in bpm sampling values through the flight. The predictions made by the model are largely based on this metric.

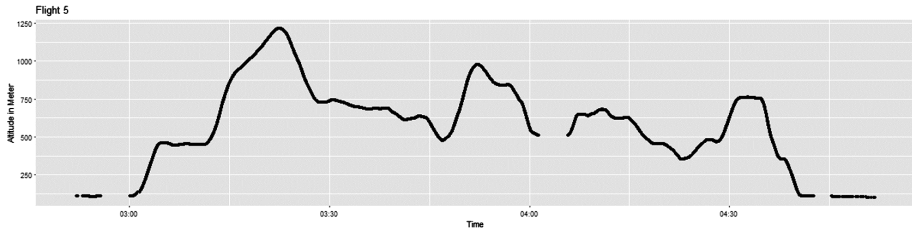


Fig. 9. Tandem 5 - Altitude in meters of the aircraft.

4 Discussion

Results regarding HR and HRV supported both hypotheses concerning the relationship between physiological parameters and roles (H1a, H1b). Indeed, as expected, mean HRV in ms was lower for the trainees when compared with instructors, and conversely for mean HR in bpm. These findings support the assumption that expertise is associated with effortless processes. This is not surprising, but not trivial either as the effects of flight dynamics (especially in aerobic flight) on physiological parameters are still largely unknown. Because aerobatic maneuvers probably require a greater deployment of physical effort when compared with regular flights, the effects associated with aerobatic flight might have prevented the effects associated with cognitive effort deployment from being observed. Fortunately, the results show that roles had a statistically significant impact on physiological parameters.

Results regarding all three variables suggested that flight phases have a significant effect on physiological parameters. The results obtained also highlighted a significant effect of the interaction of the role and phase on the RR in bpm. These results, again, were expected if we consider that different flight phases induce different levels of cognitive effort, depending on the difficulty of each phase.

Effect of flight phases can be considered as a reflection of the differences induced notably by the different procedures associated within each phase, and the variation of expertise of each pilot on these specific situations. By extension, these results raise the importance of taking into account the context of the mission and several associated external parameters, when modeling cognitive efforts and similar concepts. However, the current model of mental effort does not capture flight phases or procedures, and more generally does not take avionic parameters into account. A next step will be to link physiology-based predictions with the context of the mission. The use of avionic and contextual parameters will also allow the consolidation of the “expected good behavior” of a pilot, depending on the situation and the mission which must be performed, and hence improve the accuracy of the model. Such behavioral measures and context aware systems are deemed essential for real-world application of mental effort models and similar concepts (Elkin-Frankston et al. 2017, Bracken et al. 2016, 2017).

We argue that the model developed presented in this paper is linked with effort of mental processes, and that it can be used to quantify learning associated with a given procedure. Indeed, it can be argued that the only difference between the “role” of the

pilots (either instructor or trainee) is expertise since they were measured in tandem on similar flights. Expertise itself cannot be measured directly with physiology without context. Given the nature of the physiological data, and the support to hypotheses in a context where expertise plays a great role, it can be stated that we measured variations in physiological parameters associated with effort. Such a model is interesting because it could allow the identification of procedures which are not yet fully acquired by the trainee. If we consider the example of Tandem 5, presented in Fig. 7, the predictions made by the model do not allow the differentiation of the student from the instructor during the second part of the flight (end of flight, landing, and post-flight). This can be explained by the fact that the physiological pattern of the trainee was similar to the one of an instructor, as captured by the model. Given this information, the instructor could, if the decision of the model matches with his personal appreciation, make the decision of spending more time on other, less automated exercises, and thus individualizing the training. Such individualization lies at the heart of optimal training, especially for combat aviation population (Meland et al. 2015).

Future work will focus on the development of a feedback mechanism to the instructors and trainees, and quantification of the benefits – in terms of learning – associated with the use of this tool.

References

- Bracken, B., Palmon, N., Kellogg, L., Elkin-Frankston, S., Farry, M.A.: Cross-domain approach to designing an unobtrusive system to assess human state and predict upcoming performance deficits. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 707–711. Sage, Los Angeles (2016)
- Bracken, B., Palmon, N., Koelle, D., Elkin-Frankston, S., Farry, M.A.: Toolkit to assist researchers to more efficiently conduct experiments assessing human state. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 2032–2036, Sage, Los Angeles (2017)
- Dahlstrom, N., Nahlinder, S.: Mental workload in aircraft and simulator during basic civil aviation training. *Int. J. Aviat. Psychol.* **19**(4), 309–325 (2009)
- Dreyfus, H.L.: Intuitive, deliberative, and calculative models of expert performance. In: Zsombok, C.E., Klein, G. (eds.) *Naturalistic Decision Making*. Psychology Press, Hove (2014)
- Durantini, 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)
- Elkin-Frankston, S., Bracken, B.K., Irvin, S., Jenkins, M.: Are behavioral measures useful for detecting cognitive workload during human-computer interaction? In: Ahram, T.Z., Karwowski, W. (eds.) *Advances in The Human Side of Service Engineering*. AISC, vol. 494, pp. 127–137. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-41947-3_13
- Fornette, M.P., Darses, F., Bourgy, M.: How to improve training programs for the management of complex and unforeseen situations. In: Proceedings of the Human Factors and Ergonomics Society Europe, pp. 217–224 (2015)
- Gagnon, O., Lafond, D., Gagnon, J.F., Parizeau, M.: Comparing methods for assessing operator functional state. In: Proceedings of the 2016 IEEE Conference on Cognitive Methods in Situation Awareness and Decision Support, San Diego, CA, USA (2016)
- Kahneman, D.: Maps of bounded rationality: Psychology for behavioral economics. *Am. Econ. Rev.* **93**(5), 1449–1475 (2003)

- Kennedy, D.O., Scholey, A.B.: Glucose administration, heart rate and cognitive performance: effects of increasing mental effort. *Psychopharmacology* **149**(1), 63–71 (2000)
- Masa, J.F., Corral, J., Martin, M.J., Riesco, J.A., Sojo, A., Hernández, M., Douglas, N.J.: Assessment of thoracoabdominal bands to detect respiratory effort-related arousal. *Eur. Respir. J.* **22**(4), 661–667 (2003)
- Meland, A., Fonne, V., Wagstaff, A., Pensgaard, A.M.: Mindfulness-based mental training in a high-performance combat aviation population: a one-year intervention study and two-year follow-up. *Int. J. Aviat. Psychol.* **25**(1), 48–61 (2015)
- Nourbakhsh, N., Wang, Y., Chen, F.: GSR and blink features for cognitive load classification. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) *INTERACT 2013. LNCS*, vol. 8117, pp. 159–166. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40483-2_11
- Overbeek, T.J., van Boxtel, A., Westerink, J.H.: Respiratory sinus arrhythmia responses to cognitive tasks: effects of task factors and RSA indices. *Biol. Psychol.* **99**, 1–14 (2014)
- Roscoe, A.H.: Assessing pilot workload. why measure heart rate, HRV and respiration? *Biol. Psychol.* **34**(2–3), 259–287 (1992)
- Sanfey, A.G., Chang, L.J.: Multiple systems in decision making. *Ann. N. Y. Acad. Sci.* **1128**(1), 53–62 (2008)
- Stanton, N.A., Salmon, P.M., Walker, G.H., Baber, C., Jenkins, D.: *Human Factors Methods: A Practical Guide for Engineering and Design*. Ashgate, Farnham (2005)



Variation in Pupil Diameter by Day and Time of Day

Shannon R. Flynn¹(✉), Jacob S. Quartuccio², Ciara Sibley³, and Joseph T. Coyne³

¹ University of New Hampshire, Durham, NH, USA
srf2000@wildcats.unh.edu

² George Mason University, Fairfax, VA, USA
jquartuc@gmu.edu

³ Naval Research Laboratory, Washington, D.C., USA
{ciara.sibley, joseph.coyne}@nrl.navy.mil

Abstract. Over 60 years of prior work has shown that an individual's pupil diameter increases as the cognitive demands of a task increase. Recent work has also shown that resting pupil size is significantly correlated with an individual's working memory capacity, suggesting that between subjects variation in pupil size is important. Given the importance of both within and between variations in pupil size, the present study sought to examine the reliability of pupil diameter measurements across multiple days. A longitudinal, within subjects design was used to study pupil diameter to determine if it is possible to find stable estimates of pupil diameter across several days and across time of day. This study collected pupil data using a low-cost Gazepoint GP3 HD desktop eye tracking system. Seven participants engaged in a resting luminance change task twice per day for a total of 10 days. The participants sat in a completely dark room for two minutes prior to the start of the experiment to allow their eyes to acclimate to the darkness. They then performed the resting luminance change task, as well as two other tasks omitted from analysis. This paper presents an analysis demonstrating that there are stable estimates of pupil diameter across days, as well as across time of day. These results suggest that pupil diameter is a reliable measure within individuals. The ability to reliably capture pupil diameter using low-cost eye trackers suggests that these new low cost systems may be incorporated into a broader range of cognitive research.

Keywords: Eye tracking · Pupil diameter · Variability · Gazepoint

1 Introduction

1.1 Pupil Data as a Cognitive Measure

Pupil diameter has become a widely accepted physiological indicator of mental workload. As the amount of mental effort required for a task changes, pupil size will also change. Hess and Polt [1] were some of the first researchers to demonstrate that pupil size increases as task difficulty increases. They performed an experiment using mental arithmetic problems with increasing difficulty. Pupil diameter was recorded using a camera and

measured manually for each individual frame with a ruler. The participants typically showed an increase in pupil size which reached a maximum immediately before answering a question, and then retracted to a baseline size. The results of this study revealed that there is a strong correlation between pupil diameter and level of difficulty of a problem. Hess and Polt helped to pioneer the use of eye data as a cognitive measure in 1960. Their results help to demonstrate that pupillary response is a valuable measure for problem-solving, as well as other mental processes [1]. Soon after, many researchers in the field quickly began investigating other cognitive factors influencing pupil diameter,

Working memory is a cognitive measure which has been shown to influence pupil diameter. Kahneman and Beatty [2] quickly began to expand on the findings of Hess and Polt [1]. They were concerned with investigating the discovery that the pupil dilates while a person is listening to information, and then contracts as they report it. The task required for this study involved listening to strings of digits and reporting them back immediately. Task difficulty varied across trials with more digits being considered more difficult. The researchers determined that there was a “loading phase” in which the pupil dilates with each digit presented, and an “unloading phase” in which the pupil size would decrease with each digit being reported. They also found that the maximum pupil size obtained was correlated to the number of digits that were presented [2]. This study has influenced the utilization of a digit-span task to observe working memory fluctuations by measuring pupil diameter.

Early research in pupil diameter also identified pupil diameter differences between individuals of different cognitive ability levels. Ahern and Beatty [3] performed a study using mental multiplication problems which differed in difficulty. They found that people with higher cognitive ability, as evidenced by scores on the Scholastic Aptitude Test (SAT), performed better on the multiplication problems at every level of difficulty than those with a lower cognitive ability. What is unique about this study is that the subjects with higher cognitive ability showed smaller task-evoked pupillary dilations compared to their lesser-scoring peers [3]. The results of this allow us to identify a possible physiological measure of intelligence.

More recent research built on the findings above [3] to observe the relationship between pupil diameter of individuals during baseline and intelligence [4]. The study classified individuals as having a lower working memory capacity if they scored in the lower quartile for an OSPAN task. Pupil diameter was measured during a baseline task in which the participants looked at a dark screen. The individuals that were considered to have lower working memory capacity were determined to have significantly smaller pupil diameters than individuals in the upper quartile of working memory capacity [4]. The average pupil diameter of individuals with the higher working memory capacity was 1 mm larger than those with the lower working memory capacity. This baseline difference is considerable since pupillary increases during cognitively challenging tasks are typically less than 0.5 mm [4]. These results help to solidify using pupil diameter as a physiological measure of intelligence.

Originally, research using pupil diameter to measure cognitive effort treated it as a reporter variable. Essentially, that means it was used as one which fluctuates with cognitive processes despite having no obvious relationship to those processes [5]. Several studies in recent years have uncovered a link between activation in the locus coeruleus

(LC), the region in the brain stem responsible for production of norepinephrine, and pupil dilation [6]. The strength of the association is strong enough that many researchers now consider pupil diameter to be a means of assessing LC activation.

1.2 Other Physiological Measures

In spite of the growing research into understanding pupillary responses, there has been minimal work exploring the degree to which an individual's baseline pupil diameter and pupillary response vary across days. Pupil diameter has become a widely accepted measure of cognitive ability, yet we are still unsure how reliable these measurements are across days. Much like pupil size, other physiological responses can change with cognitive tasks, as well as show a variation from day to day [7]. For example, heart rate variability changes depending on the type of cognitive task being performed [7].

Similarly, another study sought to predict changes in performance on a cognitive task using current heart rate variability [8]. They utilized an Advanced Trail Making Test (ATMT) as their task to measure cognitive performance. To assess heart rate variability, they used an electrocardiogram, which is a less evasive than other methods of assessing cognitive performance, such as an electroencephalogram (EEG). The researchers determined that for all of the participants, a decrease in heart rate variability (as well as an increase in sympathetic and parasympathetic nerve activity), were strong contributors to a decrease in cognitive performance. They were capable of predicting performance with an 84.4% accuracy [8]. This study demonstrates a way in which heart rate variability may be similar to pupillary response, as they are both methods for determining cognitive performance or workload. Because many physiological responses of a person are similar to pupillary responses, we hypothesize that one ought to see within subject differences in pupil diameter from day to day.

1.3 Prior Dark-Adapted Research

Brown et al. [9] are some of the only prior researchers to investigate day to day variations in pupil size. This was a fairly recent study in which they looked at whether there was a pupil diameter difference when testing dark-adapted pupils. The participants were subject to different dark-adaption protocols and had their pupil size measure twice in one week, between one to seven days apart. The results indicated that they did not find any significant difference in pupil diameter. However, their participants were only measured twice, which may not be sufficient time to observe a significant difference. Also, the purpose of this study was to evaluate different dark-adaptation protocols for the preoperative assessment of refractive surgery [9]. The present study aims to build on these findings with the intention of a different use for the conclusion.

1.4 Goal

As far as we know, there have yet to be any longitudinal studies conducted investigating the overall variation of pupil diameter across more than two days. Pupil size changes as mental workload changes, but does it also change across days? There has been extensive

research on other physiological measures which may be indicators of mental workload. However, many of these methods are evasive and awkward. For example, an electrocardiogram or an electroencephalogram require a participant to be hooked up to electrodes. Methods like these restrict movement and make the participant feel uncomfortable. Not to mention, these can be very expensive tools to purchase [8]. Therefore, it is important to have a comfortable, low cost, and reliable method for measuring cognitive processes. Low-cost eye trackers seem to fit the criteria, but it has yet to be determined if pupil diameter is a reliable measure from day to day for individuals.

There are many factors that could potentially influence a person's physiological responses each day, such as amount of caffeine intake, fatigue, alertness, or sleepiness. We are interested in modeling whether there is variation of pupil diameter within subjects by day and time of day.

2 Methods

2.1 Participants

Eye tracking data were collected from 7 volunteers (4 male, 3 female) working at the Naval Research Lab. Their ages ranged from 21 to 38 ($M = 30.43$, $SD = 6.65$).

2.2 Materials

This experiment utilized a Gazepoint GP3 HD Desktop eye tracking system and pupil data were collected at 150 Hz. Each participant calibrated the system prior to the start of each experiment using the built-in calibration system from the Gazepoint control software. Data were collected on a 24 in. monitor with a 3840×2160 resolution. An Essilor digital corneal reflection pupilometer (CRP) was used to measure the interpupillary distance for each individual. The interpupillary distance was recorded for 100, 65, and 50 cm focal points.

2.3 Procedure

At the beginning of the experiment, each participant sat in the experimentation room for two minutes with the door shut, and the monitor and lights turned off. This waiting period was imposed to allow each individual's eyes to acclimate to the darkness. Immediately following, participants turned on the monitor and opened the Gazepoint software. Participants were seated approximately 60 cm from the display based on Gazepoint's guidelines. They began the calibration process for the eye tracker. The calibration was performed using the default procedure included with the Gazepoint software. This involves following a circle around the screen and pausing at 9 specific locations. Once calibration was satisfactory, the participants conducted a color change task, digit span task, and psychomotor vigilance task. The order was constant throughout the entire data collection period. However, only information from the color change task is presented below as the other tasks were not the focus of this study. Each participant performed the

experiment twice per day; one session was before lunch, and the second session was performed ~3–5 h after the first. Participants did this for a total of 10 days.

2.4 Color Change Task

The color change task was a resting luminance change task. Pupil size was captured on each individual's response to change in screen luminance. Each participant focused on a crosshair in the center of the screen. The screen started as black and remained constant for 30 s. The screen immediately changed to gray for 30 s, and then white for another 30 s. There was no transition period between the different colors. The entire process took one minute and 30 s.

Following data collection, the participants had the distance between their pupils measured. This ground truth data made it possible to convert pixels to millimeters using the pixel data generated by the Gazepoint GP3. Pupil diameter was manually converted from pixels to millimeters as the millimeter data from the Gazepoint system was found to be inaccurate. Having ground truth measurements made it possible to compare each participant's data.

2.5 Eye Tracking Data

The Gazepoint GP3 system measures left and right pupil diameter in pixels, millimeters, and the x,y position of the pupil within the system's camera. The pupil position data allowed for the computation of a third pupil diameter measurement, which used the individual's interpupillary distance recorded from the Essilor CRP at 65 cm. The distance between the eyes in pixels was divided by the individual's IPD in mm to provide a pixel to mm conversion factor. This factor was calculated for each sample and used to compute new left and right mm values for each sample from the pixel data. We converted the pupil diameter in pixels to millimeters because the millimeter data from the Gazepoint system was found to be inaccurate. Using an on-screen live measurement of pupil diameter in pixels and mm, we were able to observe that the mm data was extremely sensitive to slight shifts in distance to the eye tracker. For example, a head shift of a few centimeters closer to the screen could increase the mm data by more than a few mm, but not have any impact on the pixel data. Having ground truth measurements made it possible to compare each participant's data.

The Gazepoint system records a binary quality measure with each data point to signify whether the system considers the quality of the data point to be good or bad. We used this as a general filtering method and removed any data the system considered to be bad.

2.6 Analysis

All of the analysis was conducted in R [10]. In past studies [11], left and right pupil size have been highly correlated ($R = .90$). Therefore, we only used data from the left pupil for this analysis.

The median pupil diameter of each person at each session was calculated. Medians were taken for the entire session, as well as for each individual background color. Medians were used to help reduce the effects of outliers in the data. There were a few missing time points – therefore, the mice package [12] was used to impute data. The predictive mean matching method was used and only used the first data set in the subsequent analysis since there were only 14 missing sessions (out of 140) due to missing data. Coefficient alpha (Cronbach) was then calculated to estimate the reliability of the eye data. A visual display was also created to show the median values of each participant which was calculated using ggplot2 [13].

3 Results

Coefficient alpha was calculated for the data described above. The analysis revealed that the data produced highly reliable results. With the seven participants across 20 measurement sessions (10 days, twice per day), the overall reliability across the color change task is $\alpha = 0.98$. The results are also highly reliable for each individual background color. Black was shown to have an α of 0.98, Gray an α of 0.99, and White an α of 0.98 (Figs. 1, 2, 3 and 4).

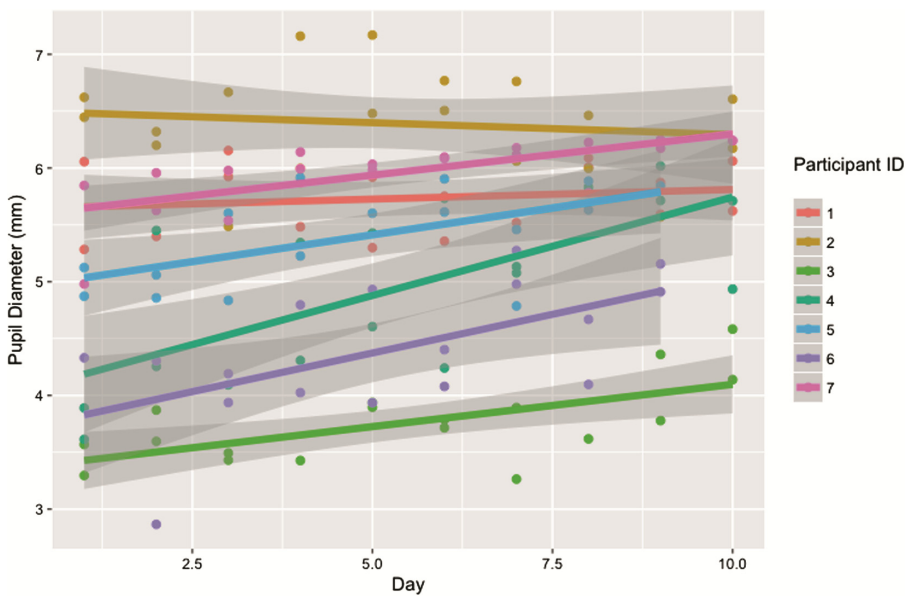


Fig. 1. Median pupil diameter (mm) for each session each day, averaged across all three background colors of the Color Change Task. (Color figure online)

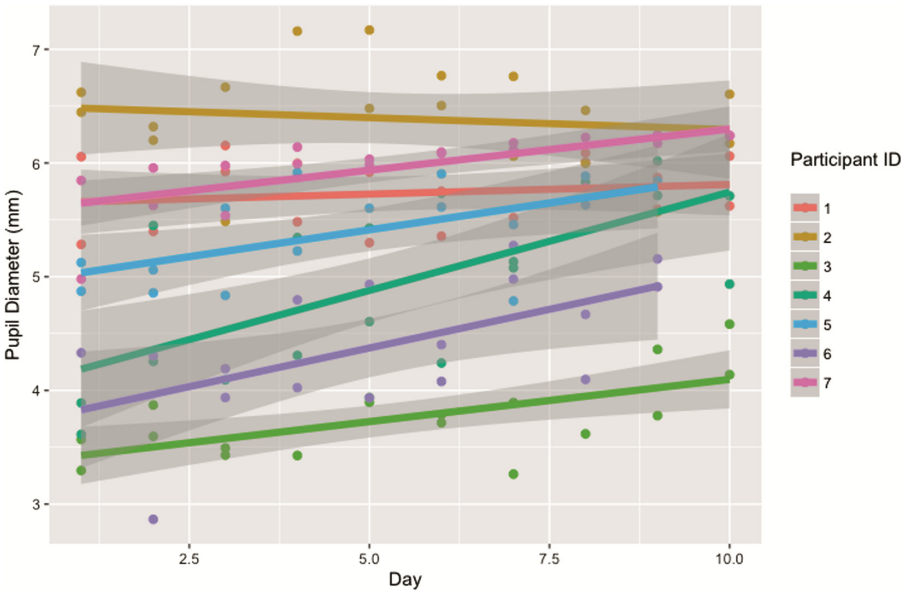


Fig. 2. Median pupil diameter (mm) for each session each day while looking at a black screen, by participant. (Color figure online)

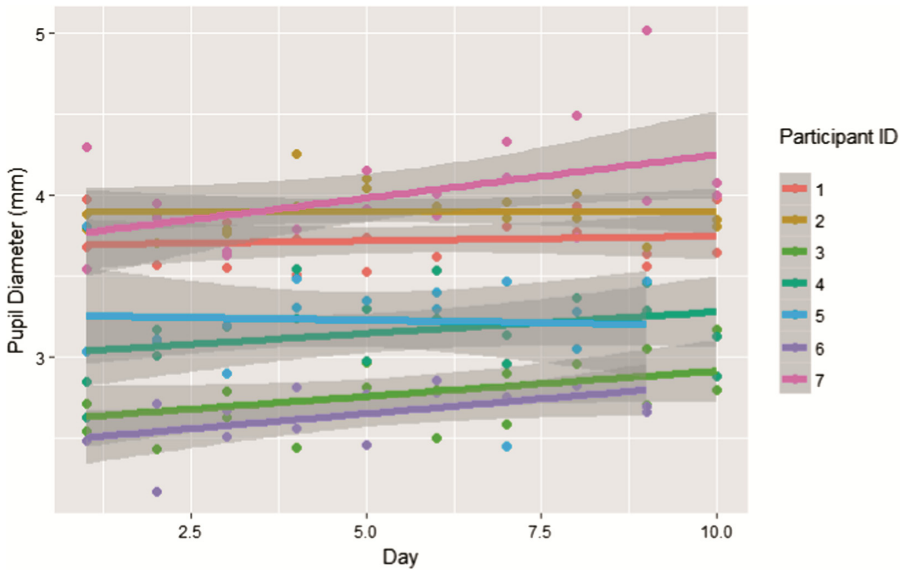


Fig. 3. Median pupil diameter (mm) for each session each day while looking at a gray screen, by participant. (Color figure online)

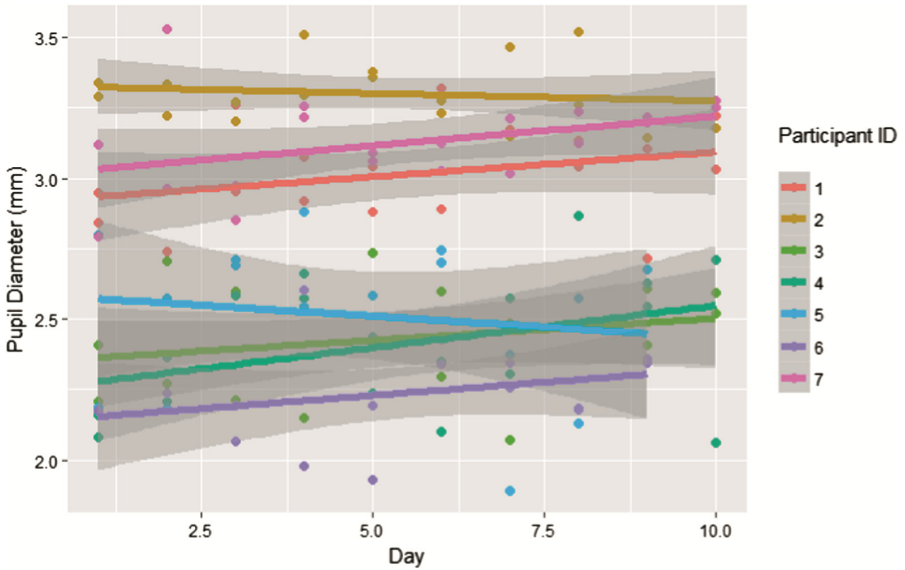


Fig. 4. Median pupil diameter (mm) for each session each day while looking at a white screen, by participant.

A two sample t-test was also performed to compare pupil diameter from AM sessions and PM sessions. There was no evidence of a significant difference in pupil diameter across time of day ($t(121.12) = -0.687, p = .4933$).

These results are highly reliable, except there was a slight abnormality observed in the data; the median pupil diameter appears to rise across days, which can be seen in the figures above. We are unsure of the exact reason for this second effect, and it could require further investigation to determine the underlying causes.

4 Discussion

The goal of this study was to determine whether pupil diameter is a constant measure across days for individuals. Using coefficient alpha, this study revealed pupil size is a highly reliable measure and it does not significantly change across days. Also, through a two sample t-test, it was shown that pupil diameter does not significantly differ across time of day.

Future research should attempt to replicate this procedure and possibly use a longer longitudinal study. It may be beneficial to see if other studies find the same result of median pupil size rising across days, and further investigate this abnormality. Additionally, it would be useful to explore if there are any factors that may contribute to pupil size variation within subjects. For example, a person who has more caffeine in the morning may have larger variation by time of day than someone who does not intake caffeine. This could be explored using a self-report questionnaire before each session to determine a person's fatigue, sleepiness, amount of stress, etc. This could also be

manipulated through a design in which one controls for variables, such as drugs or amount of sleep. Altering a person's state from one day to the next may show different results than what were revealed with this study.

Future research should also investigate day to day variability across a variety of cognitive tasks, such as memory, attention, arithmetic, and vigilance tasks. Since research has shown that cognitive demands influence pupil size, it would be useful to see how reliably pupil data can be used to measure tasks which require mental effort across multiple days. We collected data on two cognitive tasks (digit span and psychomotor vigilance) but did not analyze it for the use of this study.

Overall, the stability of pupil diameter, as assessed via a low-cost eye tracker, suggests that the equipment can be sensitive not only to changes within an individual but also able to differentiate across individuals. However, it is important to note that these results were obtained using pixel data that was converted to mm based upon an individual's interpupillary distance and not the system's provided mm data. Although devices to measure interpupillary distance are also inexpensive it does add an extra step in data collection and processing. It also suggests that the mm data obtained directly from the Gazepoint software may not be reliable. Therefore, future studies might consider using the same process of converting pixel data into mm data manually, rather than performing analyses using the given mm data.

The findings of this study ought to be encouraging. The ability to reliably capture pupil diameter using low-cost eye trackers suggests that these new low cost systems may be incorporated into a broader range of cognitive research. Pupil diameter has become a widely used physiological measure of cognitive information, such as intelligence [3], and working memory capacity fluctuations [2]. The ability to reliably capture this data using low-cost systems, rather than something more expensive and uncomfortable, such as an electroencephalography or electrocardiogram, should persuade more researchers to opt for this low cost alternative.

References

1. Hess, E.H., Polt, J.M.: Pupil size in relation to mental activity during simple problem-solving. *Science* **143**(3611), 1190–1192 (1964)
2. Kahneman, D., Beatty, J.: Pupil diameter and load on memory. *Science* **154**, 1583–1585 (1966)
3. Ahern, S., Beatty, J.: Pupillary responses during information processing vary with scholastic aptitude test scores. *Science* **205**(21), 1289–1292 (1979)
4. Tsukahara, J.S., Harrison, T.L., Engle, R.W.: The relationship between baseline pupil size and intelligence. *Cogn. Psychol.* **91**, 109–123 (2016)
5. Beatty, J., Lucero-Wagoner, B.: The pupillary system. In: *Handbook of Psychophysiology*, 2nd edn, pp. 142–162. Cambridge University Press, New York (2000)
6. Aston-Jones, G., Cohen, J.D.: An integrative theory of locus coeruleus- norepinephrine function: adaptive gain and optimal performance. *Neuroscience* **28**, 403–450 (2005)
7. Luque-Casado, A., Zabala, M., Morales, E., Mateo-March, M., Sanabria, D.: Cognitive performance and heart rate variability: the influence of fitness level. *PLoS ONE* **8**(2), 1–9 (2013)

8. Tsunoda, K., Chiba, A., Yoshida, K., Watanabe, T., Mizuno, O.: Predicting changes in cognitive performance using heart rate variability. *Inst. Electron. Inf. Commun. Eng.* **E100-D**, 2411–2419 (2017)
9. Brown, S.M., Khanani, A.M., Xu, K.T.: Day to day variability of the dark-adapted pupil diameter. *J. Cataract Refract. Surg.* **30**, 639–644 (2004)
10. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2014)
11. Foroughi, C.K., Coyne, J.T., Sibley, C., Olson, T., Moclair, C., Brown, N.: Pupil dilation and task adaptation. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) *AC 2017. LNCS (LNAI)*, vol. 10284, pp. 304–311. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58628-1_24
12. van Buuren, S., Groothuis-Oudshoorn, K.: MICE: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**(3), 1–67 (2011)
13. Wickham, H.: *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York (2009). <https://doi.org/10.1007/978-3-319-24277-4>



Computer-Based Neuropsychological Assessment: A Validation of Structured Examination of Executive Functions and Emotion

Gilberto Galindo-Aldana¹(✉), Victoria Meza-Kubo², Gustavo Castillo-Medina²,
Israel Ledesma-Amaya¹, Javier Galarza-Del-Angel¹, Alfredo Padilla-López¹,
and Alberto L. Morán²

¹ Engineering and Business School, Guadalupe Victoria, Research Group of Mental Health,
Society and Profession, Universidad Autónoma de Baja California, Mexicali, México
gilberto.galindo.aldana@uabc.edu.mx

² Faculty of Sciences, Research Group of Technology for Intelligent Environments,
Universidad Autónoma de Baja California, Ensenada, México

Abstract. An increase in the use of Computer-Based Neuropsychological Assessment tools (CBNA) has approached clinical neuropsychology appliance. In clinical diagnosis practice it is strongly needed to acquire precise data which often presents a challenge for clinicians and neuroscientists. Procedures for validation of methods in clinical neuropsychology are reliable when results between clinical and control samples are expected and observed different, by using paper-based and computer-based methods. The aim of the present study is to describe the validation procedures of a CBNA tool in a sample of control and clinical participants. The method consisted in comparing 35 control adolescents with 33 clinically referred pairs. A CBNA composed by two neuropsychological assessment tests for measuring effect of emotions on executive functions, was administered to each participant. Results showed differences between groups, observed in performance over the tasks. It was concluded that CBNA gives accurately results that otherwise could not be acquired by conventional paper-based methods, reducing errors of tests administration and application costs, as well as conserving reliability.

Keywords: Computer-based · Assessment · Neuropsychology

1 Introduction

There is a significant increase in the use of Computer-Based Neuropsychological Assessment (CBNA), in clinical diagnosis practice in different specialties, such as neuroscience and cognition dyslexia diagnosis [1, 2], sports-related concussion [3, 4], human computer biological signals interaction [5], neurologic patients [4]. Since CBNA offers several advantages when applied to different clinical disorders in neuroscience, it is possible to accurately control variables for measuring cognitive functions, such as reaction time, correct and incorrect responses, error types, and direct stimuli administration. CBNA brings a great advantage for clinicians to dynamically manage assessment

sessions, reducing the risk of errors related to sequence of tasks during an evaluation. Furthermore, it significantly simplifies quantitative analysis of the outcome results, by automatically calculating the desired data with high definition and data grouping according to neuropsychological clinical models previously established. Administration of assessment procedures in neuropsychology in general requires at least one measure which needs to be computed, scored or counted [6].

Other important advantages are identified when using CBNA instruments, such as reliability and efficient use of available resources. When assessing executive functions there are specific considerations that a clinician must take care about, minimal deflections of procedures during administration of the tests, could lead to a complete discard of results. For example, an assessment instrument called Wisconsin Card Sorting Test, has shown good reliability when measuring executive functions. A study by Tien et al. [7] which compared the performance between paper-based (typical cards variant of the test and a computer-based version in psychiatric participants, showed more reliable administration and accuracy in data acquisition and scoring, when using the computer version.

Adolescence is behavioral and physiological, including hormonal changes age and stage of life that requires specific assessment methods. Puberty supposes a cascade of changes over the endocrinal system that involves brain and psychological processes related to affect regulation [8]. The mentioned changes often require accurate clinical assessment procedures for taking decisions about improving development. Likewise, cognitive competence in adolescence includes the ability to reason effectively, problem solve, think abstractly and reflect, and plan for the future. Despite their rapidly developing capacity for higher-level thinking, most adolescents still need guidance from adults to develop their potential for rational decision making [9]. According to Giedd [10] in a longitudinal study review related to teenager brain assisted by neuroimaging, investigations report that at ages from 7–29 years, adolescents respond to reward particularly in regions corresponding to the nucleus accumbens, which was equivalent to that found in adults, but adolescent orbitofrontal activity was similar to the equivalent area in children, fact that may be associated to a lack of maturity of rational decision making functioning. Similarly, maturation of prefrontal cortex that regulates judgment, caution, and appropriate behavior is a relatively late in adolescence, early adult [11]. Other studies indicate that the level of maturation of intelligence activity is related to the trajectory of cortical thickening during its development through childhood and adolescence, primarily in frontal regions [12]. Development of frontal cortex through adolescence period, has relevant changes that bias maturity of executive functions. Executive function developmental process, ends between the second and third decades of life, specifically in a critical period between 12–15 years old where cognitive functions improvement such as, mental flexibility and sequencing, visuospatial and sequential planning, verbal and visuospatial working memory, and risk and benefit mental processing [13], functions that have an important role in self-control, coordination of thoughts and adaptive behavior to diverse circumstances of lifespan [14]. According to research results [15] assessment of executive functions during adolescence, is critical as a dominant factor among other non-cognitive skills related to educational performance, health behavior and delinquency or substance abuse.

There are instruments such as the Self-descriptive Adolescents Inventory [16], dedicated to evaluate personality and social skills in adolescents, the procedures of application of those instruments are complex. Limitations of neuropsychological executive functions instruments often are related to a lack of well-standardized, developmental assessment techniques [17]. When geographical conditions represent a challenge to resolve, as often occurs in rural areas, as well as vulnerable conditions where clinical services are difficult to approach, remote administration of neuropsychological assessment tests by minor requirements or internet facilitates these clinical procedures [3]. Previous built CBNA as the NeuroCogFX [4], had included executive function tasks for assessing short term memory, working memory, psychomotor speed, selective attention, verbal and figural memory and verbal fluency, which resulted in a brief 25 min structured battery supported on statistical reliability and standards.

The study of the impact of computer-based assessment tools has been extended even to academic and educational areas [18]. Results in the field of academic assessment, had demonstrated that after familiarization period, computer-based assessment performance, and acceptance is an alternative to pen-and-paper theoretical practical examinations.

It has been reported that computer familiarity is related to performance on computerized neurocognitive assessment [19]. Particularly adolescents had recently increased their screen-time spent, which is reported as a factor of behavior modification [20], but can also be taken as an advantage for improving clinical assessment methods. The aim of the present study is to describe the validation procedures of a CBNA tool in a sample of control and clinical adolescent participants.

2 Method

This study starts from previously validated and elsewhere published clinical assessment tasks [21, 22], which are commonly used by neuropsychologists for executive functions and emotion interaction diagnosis, for different cognitive disorders and syndromes.

2.1 Instruments and Procedures

Montreal Cognitive Assessment (MoCA) [23] was used as a screening test to measure global cognitive state. The CBNA software was developed by requirements from a neuroscience and cognition laboratory. It is constituted by an Emotionally Interfered Working Memory task (EIWM), composed by 36 emotionally loaded BMP format images, divided in three groups (pleasant, unpleasant, and neutral). And a modified Iowa Gambling Task (mIGT) characterized by 4 card decks configured according to the neuropsychological assessment model, in which the participant can lose or gain virtual money.

Required modules were developed using JAVA, and were integrated in a unique interface from which can be evocated (Fig. 1). All logic and flow rules are embedded in the code, as well as the acquired results of the task, which are exported to a file. Options for the identification of patient options were included in order to keep a record of the

assessed participants. The results of the application of the CBNA record particular indicators of the neuropsychological task features, including type of visual stimuli (pleasant, unpleasant, neutral, or facial emotional expression). For the mIGT, the software is able to record the cumulative score obtained by the participant, by considering the losses or gains. Visual stimuli can be presented in any size monitor, however, for clinical purposes a 19" screen is recommended to ensure the emotional effect of the visual stimuli. Gambling task responses can be obtained by means of the mouse, or a touch screen. EIWM require the use of an external numeric keyboard, marked with a red spot over the number 4, and a green spot over the number 6 to acquire accuracy of response and reaction time. Indicators within reaction times are exported automatically to a predefined folder in the system, in CSV (comma separated values) format, and it is labeled by the task name, participant's name, and application date. Visual stimuli images are presented in 1024×768 pixels (Fig. 1).

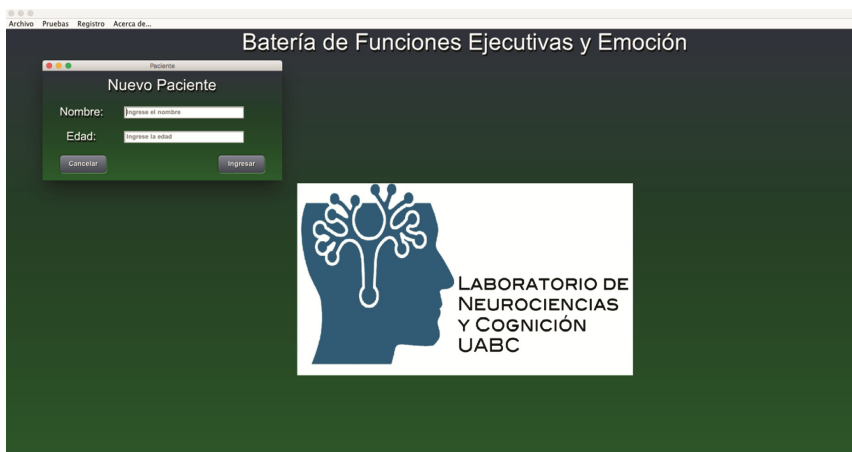


Fig. 1. CBNA system screen sample illustrating initial display for participant registration. (Color figure online)

EIWM is constituted of three blocks of 21 equally divided pleasant, unpleasant and neutral visual stimuli from the International Affective Picture System. The configuration of EIWM task was measured using previous designed and validated method [22, 24] facial images, objects, and scenes having one of three different emotional valences retrieved from the International Affective Picture System (IAPS) [25]. From this database, 42 pictures were selected, and divided into three groups: 14 associated with positive emotional states, 14 associated with negative emotional states, and 14 considered neutral. The images were presented in pairs with 4 s delay between the first and second pictures, the pairs of images could be the same, meaning that the second picture was not different from the first one, or different, meaning that the picture could have one or more features that are different from the first picture (50% of the pairs were the same and 50% were different). Each of the pairs of pictures differed from other pairs in the emotional valence being shown, in addition to being the same or different, and were presented in

a pseudorandom order. The participants were asked to maintain in memory the first image of the pair, and compare the second image with the stored image. Participants were asked to press a green button on a keyboard if the pictures were the same and a red button if the pictures were different, a lag of 10 s was given to respond. If no response was given during that time period, the next pair of pictures was presented. Accuracy of response (AR) as measured by the number of correct responses made and reaction times (RT) were measured in milliseconds for each participant. The assessments were individually conducted in a quiet room where the visual stimuli were shown on a 19 inch color monitor at a distance of 40 cm from participant's face.

mIGT was a 50 card sequence applied, each card was previously configured and controlled by the person who applied the protocol. The configuration was based on previous developed test procedures [26]. Each participant received a \$2000 (virtual Mexican Peso) credit, and had to choose a card at a time from four possible decks A, B, C, and D. Each card could sum or rest money to the participant's credit. In this variant there was no time limit, and the participant was allowed to change from card deck as many times as he or she wanted, and that the goal of the task was to acquire the largest possible money amount. A research assistant remained close to the participant to give support and resolve doubts during the instructions part of the study, the participant was free to ask questions to the assistant.

2.2 Participants

The sample was a total of 35 control participants, 14.1 years old mean age ($SD = 1.6$), 53% male, 47% female, 22.6 MoCA Score, and 33 clinically referred volunteer participants, 13.47 years old mean age ($SD = 1.23$), 63% male, 37% female, 17.9 MoCA Score, whose consultation reason were mainly impulsive behavior. All right handed, volunteer participants signed an informed consent. The study was considered as non-risk, considered all human rights and Helsinki Accord criteria, and was properly approved by the Bioethics Committee of the Medicine and Psychology Faculty of Autonomous University of Baja California.

2.3 Data Analysis

Acquired data was analyzed by two methods, (a) gold-standard comparison with conventional paper based task, for the mIGT, and (b) a two independent samples t-student test between control and experimental participants for both applied tasks: mIGT and EIWM.

2.4 Results

Paper-based mIGT showed statistical difference between groups ($t = -11.14$, $DF = 308$, $p > .001$) means are shown in Fig. 2. A similar result was found for EIWM reaction time for pleasant stimuli ($t = 23.1$, $DF = 308$, $p > .001$), as well as for unpleasant type of stimuli ($t = 19.4$, $DF = 308$, $p > .001$) between groups. An effect of longer reaction times was observed in means of response for each type of stimuli (Fig. 3).

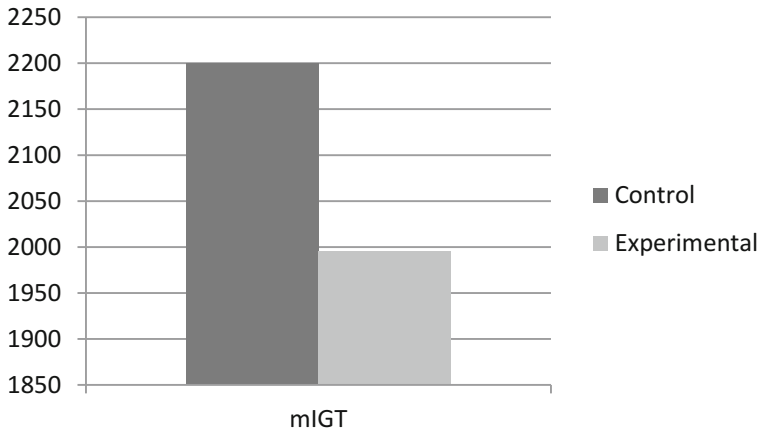


Fig. 2. Means of mIGT scores obtained between control and experimental groups.

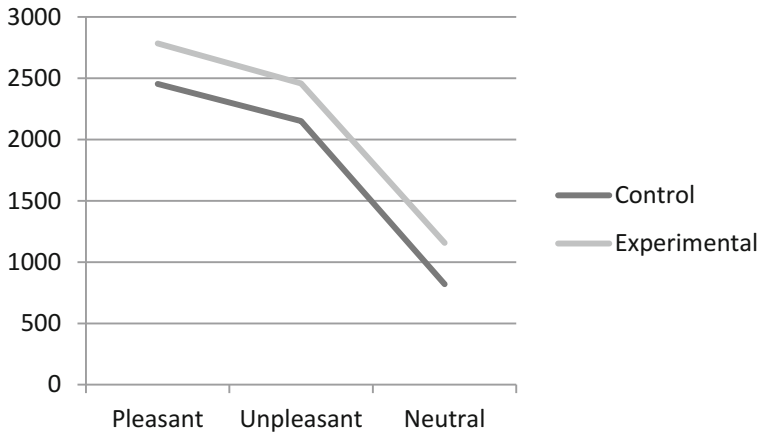


Fig. 3. Means of reaction times in milliseconds observed between control and experimental participants resolving the EIWM task.

Otherwise, accuracy of response also presents differences between types of stimuli (Fig. 4) observed in the mean of number of correct responses. This suggests a lower performance of the experimental group compared to their control pairs, and an effect of reduction of accuracy in contrast with reaction time.

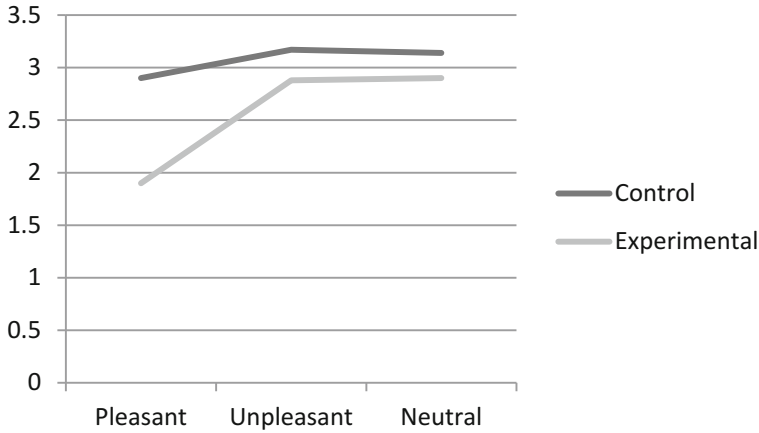


Fig. 4. Mean of number of correct responses observed between control and experimental participants resolving the EIWM task.

3 Discussion and Conclusions

CBNA showed the same accurate results when applied in clinical conditions as paper-based instruments, particularly mIGT in this study. Significant differences are found between groups, mainly due to the clinical conditions of participants. Furthermore, CBNA is highly sensitive when measuring reaction times, which is highly difficult to practice with paper-based procedures.

Studies have shown that, gathering of clinical data through digital tests is relevant to detect neuropsychological impairment. It has been observed [27], that scanning of the cognitive function test (CogState) offers, a clear evidence of the neuropsychological alterations related to HIV. In the same way, they reported that the Cogstate instrument has clinical sensitivity and specificity to detect cognitive impairment in these patients. Likewise, with similar procedures [28] another finding was that the use of CBNA in patients with concussions as a result of sports practice, can be useful to detect subtle negative changes in the cognitive process, allowing the neuropsychologist to determine how long patients should wait before starting their training, and therefore, continue participating in sports competitions. Other studies had reported significant differences in the cognitive performance of three groups, two of them, were patients with Parkinson's disease, with and without mild cognitive impairment, in contrast to a control group, using a technique virtual evaluation and a classic pencil and paper method [29]. Researchers observed that patients with Parkinson's disease with normal cognitive performance recorded by classical neuropsychological technique and the virtual executive test, show differences, even though in the classical evaluation this group score as a normal; therefore, these results show that use of virtual neuropsychological techniques had shown sensitivity to detect cognitive changes as a predictor of major neurocognitive impairment.

The advantages of using CBNA, reside in the efficacy reached through diverse aspects, like time reduced procedures, application and precision evaluation of tests, integrating some clinical relevant performance characteristics that participants exhibit during test application as the test's stimuli control presentation, which could derive in an increase of test reliability, thus CBNA allows the clinician, focalize his/her attention in treatment; results from other studies [3] suggests the use of CBNA decreases the influence of examiners. Another benefit as also mentioned by other researchers [28] CBNA could be applied at the same time to certain amount of people, without requiring that more than one neuropsychologist to check the test procedures.

One strength of our study was to leave a precedent, about the relevance of CBNA and the study of digital assessment in a Hispanic speaker population with the objective to create in the future new tools that help and enhance assessment neuropsychology techniques that allow neuropsychologists to make early diagnosis, in addition to the repercussion in the creation of an intervention plan based in virtual environments.

In regards to study limitations they were as follows, the narrowed use of our study sample to observe difference between performance in a Paper-based mIGT group and CBNA groups in order to generalize our results over a population, which allows to accomplish an external validity as a psychometric criteria. Another point is despite the control of variables, the lack of familiarity with technology in our sample could interfere with task performance, according to previous studies [19, 30] individuals with greater computer experience perform better on computer-based assessments than those with less computer experience.

Concluding, the use of CBNA have to take into account psychometric criteria, mainly focusing on two kinds of validity, an ecological one and other of localization data. The first related to prediction of a particular construct with regard to real life abilities and execution of activities of daily living, and the other, associated to test accurately focal lesions. Another relevant point is the characteristic of CBNA referred to clinical aspects, such as sensibility test property to detect subtle neuropsychological abnormalities and specificity test characteristic, or the fact to differentiate neuropsychological deficits between patients. Furthermore, it is important to mention the use of normative data, related to demographic patient's provenience. It would help to avoid diagnosis errors in individuals who do not have impairment [31]. Another related idea is the future of CBNA as regards the standardization procedures which have to be updated according to emerging technology, like gadgets, taking into account the use of apps through portable devices it has become an advantage nowadays [32]. It should be noted, the study of new virtual environments and how they would improve the detection of subtle changes in cognitive abilities until now, they are not precisely detected with paper-pencil neuropsychological tests [6].

Recent research suggest that new skills are arising from activity performance of actual life, such as information and communication technologies [33], or writing [34]. Such skills may no longer be assessed by conventional manual or paper methods, because of technical and accuracy limitations.

This particular CBNA battery offers an alternative for cognition-emotion interaction in human cognitive processing, well known as hot executive function [35]. Special attention to hot executive functions assessing in adolescents, by using the Cambridge

gambling task offers an important predictor of developmental outcomes related to emotional problems during this age [36], but may be related to other psychological disorders like obsessive-compulsive [37], and autism spectrum disorder [38].

Future work may suggest ethical considerations about using CBNA instruments. Digital availability of clinical instruments could lead to inappropriate use of these type of tools, which at the same time may lead to misunderstanding of results, by non-specialists users. For this, and other reasons, previous studies [7] indicate that computer versions of CBNA tests, should not be the substitute for the human clinicians, but a reliable tool that could carry out mechanistic processes of tests administrations, such as presentation of stimuli or scoring.

Considering these ideas, we thought that neuropsychologists have to adapt their clinical procedures, according to the available technology up to now, as a response of the constant technology advances related to the digital era.

References

1. Zygouris, N.C., Vlachos, F., Dadaliaris, A.N., Oikonomou, P., Stamoulis, G.I., Vavougiou, D., et al.: A neuropsychological approach of developmental dyscalculia and a screening test via a web application. *Int. J. Eng. Pedagog.* **7**(4), 51–65 (2017)
2. Zygouris, S., Tsolaki, M.: Computerized cognitive testing for older adults: a review. *Am. J. Alzheimer's Dis. Other Dementias* **30**(1), 13–28 (2015)
3. Schatz, P., Browndyke, J.: Applications of computer-based neuropsychological assessment. *J. Head Trauma Rehabil.* **17**(5), 395 (2002)
4. Fliessbach, K., Hoppe, C., Schlegel, U., Elger, C.E., Helmstaedter, C.: NeuroCogFX—a computer-based neuropsychological assessment battery for the follow-up examination of neurological patients. *Fortschr. Neurol. Psychiatr.* **74**(11), 643–650 (2006)
5. Meza-Kubo, V., Morán, A.L., Carrillo, I., Galindo, G., García-Canseco, E.: Assessing the user experience of older adults using a neural network trained to recognize emotions from brain signals. *J. Biomed. Inf.* **62**, 202–209 (2016)
6. Parsey, C.M., Schmitter-Edgecombe, M.: Applications of technology in neuropsychological assessment. *Clin. Neuropsychol.* **27**(8), 1328–1361 (2013)
7. Tien, A.Y., Spevack, T.V., Jones, D.W., Pearlson, G.D., Schlaepfer, T.E., Strauss, M.E.: Computerized Wisconsin card sorting test: comparison with manual administration. *Kaohsiung J. Med. Sci.* **12**(8), 479–485 (1996)
8. Cameron, J.L.: Interrelationships between hormones, behavior, and affect during adolescence: understanding hormonal, physical, and brain changes occurring in association with pubertal activation of the reproductive axis. Introduction to part III. *Ann. New York Acad. Sci.* **1021**, 110–123 (2004)
9. American Psychological Association: *Developing Adolescents: A Reference for Professionals*. APA, Washington DC (2002)
10. Giedd, J.N.: The teen brain: insights from neuroimaging. *J. Adolesc. Health: Off. Publ. Soc. Adolesc. Med.* **42**(4), 335–343 (2008)
11. McAnarney, E.: Adolescent brain development: forging new links? *J. Adolesc. Health* **42**, 321–323 (2008)
12. Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., et al.: Intellectual ability and cortical development in children and adolescents. *Nature* **440**(7084), 676–679 (2006)

13. Flores, J., Ostrosky, F.: Desarrollo neuropsicológico de los lóbulos frontales y funciones ejecutivas. Moderno, M. (ed.) México (2012)
14. Luria, A.R.: Higher Cortical Functions in Man. Springer, Heidelberg (1983). <https://doi.org/10.1007/978-1-4615-8579-4>. Barcelona Fontanella
15. Coneus, K., Laucht, M.: The effect of early noncognitive skills on social outcomes in adolescence. *Educ. Econ.* **22**(2), 112–140 (2014)
16. Gómez-Maqueo, B.: Inventario Autodescriptivo del Adolescente: Manual Moderno (2012)
17. Anderson, V.A., Anderson, P., Northam, E., Jacobs, R., Catroppa, C.: Development of executive functions through late childhood and adolescence in an Australian sample. *Dev. Neuropsychol.* **20**(1), 385–406 (2001)
18. Guimarães, B., Ribeiro, J., Cruz, B., Ferreira, A., Alves, H., Cruz-Correia, R., et al.: Performance equivalency between computer-based and traditional pen-and-paper assessment: a case study in clinical anatomy. *Anat. Sci. Educ.* **11**, 124–136 (2017)
19. Iverson, G.L., Brooks, B.L., Ashton, V.L., Johnson, L.G., Gualtieri, C.T.: Does familiarity with computers affect computerized neuropsychological test performance? *J. Clin. Exp. Neuropsychol.* **31**(5), 594–604 (2009)
20. Kurek, A., Jose, P.E., Stuart, J.: Discovering unique profiles of adolescent information and communication technology (ICT) use: are ICT use preferences associated with identity and behaviour development? *Cyberpsychology* **11**(4), 1–18 (2017)
21. Bechara, A.: The role of emotion in decision-making: evidence from neurological patients with orbitofrontal damage. *Brain Cogn.* **55**(1), 30–40 (2004)
22. Galindo, G., Fraga, M., Machinskaya, R., Solovieva, Y., Mangan, P.: Effect of emotionally valenced stimuli on working memory performance. *Psychol. Neurosci.* **8**(3), 333–340 (2015)
23. Ciesielska, N., Sokołowski, R., Mazur, E., Podhorecka, M., Polak-Szabela, A., Kędziora-Kornatowska, K.: Is the montreal cognitive assessment (MoCA) test better suited than the mini-mental state examination (MMSE) in mild cognitive impairment (MCI) detection among people aged over 60? Meta-analysis. *Psychiatr. Pol.* **50**(5), 1039–1052 (2016)
24. Rozovskaya, R., Machinskaya, R., Pechenkova, E.: The influence of emotional coloring of images on visual working memory in adults and adolescents. *Hum. Physiol.* **42**(1), 69–78 (2016)
25. Lang, P., Bradley, M., Cuthbert, B.: International affective picture system (IAPS): affective ratings of pictures and instruction manual. University of Florida, Gainesville, FL (2008)
26. Bechara, A., Damasio, A.R., Damasio, H., Anderson, S.W.: Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* **50**(1–3), 7–15 (1994)
27. Blanch, J., Muñoz-Moreno, J.A., Reverte, R., Ayuso-Mateos, J.L.: Neurocognitive deficits in patients with human immunodeficiency virus infection. *Handb. Clin. Neurol.* **106**, 589–605 (2012)
28. Segalowitz, S.J., Mahaney, P., Santesso, D.L., MacGregor, L., Dywan, J., Willer, B.: Retest reliability in adolescents of a computerized neuropsychological battery used to assess recovery from concussion. *Neurorehabilitation* **22**(3), 243–251 (2007)
29. Cipresso, P., Albani, G., Serino, S., Pedroli, E., Pallavicini, F., Mauro, A., et al.: Virtual multiple errands test (VMET): a virtual reality-based tool to detect early executive functions deficit in Parkinson’s disease. *Front. Behav. Neurosci.* **8**, 405 (2014)
30. Tun, P.A., Lachman, M.E.: The association between computer use and cognition across adulthood: use it so you won’t lose it? *Psychol. Aging* **25**(3), 560–568 (2010)
31. Carey, C.L., Woods, S.P., Rippeth, J.D., Gonzalez, R., Moore, D.J., Marcotte, T.D., et al.: Initial validation of a screening battery for the detection of HIV-associated cognitive impairment. *Clin. Neuropsychol.* **18**(2), 234–248 (2004)

32. Parsons, T.D.: Neuropsychological assessment using virtual environments: enhanced assessment technology for improved ecological validity. In: Brahnam, S., Jain, L.C. (eds.) *Advanced Computational Intelligence Paradigms in Healthcare 6 Virtual Reality in Psychotherapy, Rehabilitation, and Assessment*, pp. 271–289. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-17824-5_13
33. Engelhardt, L., Goldhammer, F., Naumann, J., Frey, A.: Experimental validation strategies for heterogeneous computer-based assessment items. *Comput. Hum. Behav.* **76**, 683–692 (2017)
34. White, S., Kim, Y.Y., Chen, J., Liu, F., National Center for Education Statistics: Performance of Fourth-Grade Students in the 2012 NAEP Computer-Based Writing Pilot Assessment: Scores, Text Length, and Use of Editing Tools. Working Paper Series. NCES 2015-119. National Center for Education Statistics (2015)
35. MacKenzie, L.E., Patterson, V.C., Zwicker, A., Drobinin, V., Fisher, H.L., Abidi, S., et al.: Hot and cold executive functions in youth with psychotic symptoms. *Psychol. Med.* **47**(16), 2844–2853 (2017)
36. Poon, K.: Hot and cool executive functions in adolescence: development and contributions to important developmental outcomes. *Front. Psychol.* **8**, 2311 (2018)
37. Hybel, K.A., Mortensen, E.L., Lambek, R., Thastum, M., Thomsen, P.H.: Cool and hot aspects of executive function in childhood obsessive-compulsive disorder. *J. Abnorm. Child Psychol.* **45**(6), 1195–1205 (2017)
38. Kouklari, E.-C., Tsermentseli, S., Monks, C.P.: Hot and cool executive function in children and adolescents with autism spectrum disorder: cross-sectional developmental trajectories. *Child Neuropsychol.: J. Normal Abnorm. Dev. Child. Adolesc.* 1–27 (2017)



The Mapping Between Hand Motion States Induced by Arm Operation and Surface Electromyography

Tingting Hou, Chen Qian, Yanyu Lu, and Shan Fu^(✉)

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

Abstract. The mental workload has been playing more and more important role in air transportation industry. And the level of mental workload has apparent link to the human operational performance. To explore the human operational performance, relationship between the states of muscles of the forearm and Surface Electromyography (sEMG) signals induced by specific motion modes should be studied first. In this paper, the flexor carpi ulnaris, flexor carpi radialis, brachioradialis, palmaris longus and biceps brachii of the right forearm are selected as the source of the sEMG signals according to the anatomy. The sEMG signals of these muscles were obtained in a specific and real experimental environment. A method of binary coding was applied to deal with the sEMG. The result of experiment shows that sEMG signals have strong ability of recognizing different hand movements. But only using the parameter of time domain can we hardly distinguish hand gestures.

Keywords: Mental workload · Human performance · Surface EMG
Binary coding

1 Introduction

With the increasing technical advancement and system complexity, modern systems being operated in dynamically environment require constant human supervision although the amount of direct human operation has been reduced considerably. As the matter of facts, the relative share of human workloads, the mental workload rather than physical workload, is increased across all modern applications, especially in the air transportation industry. More and more evidences showed that human errors have become the main reasons to result in aviation disaster [1].

In order to make sure the safety, comfort, and continued productive efficiency of the operator, a reasonable goal is to regulate mental workload so that they neither underload nor overload an individual. The evaluation of mental workload is also an important aspect of human factors research. The operator who bear abnormal workload would hardly maintain performance at required levels [2]. And it is a widely accepted assumption that the level of mental workload has apparent link to the human operational performance, especially the potential occurrence of unsafe operation of the system.

It was also realized that the performance of a human operator could be affected by the complexity of a particular task as well as various factors of the human operator during the operation of the system such as fatigue, stress, mental workload, attention deficit, and executive function, among others, which leads to errors, accidents, or even disasters. Moreover, the external characteristics of operator's workload can be measured using physiological parameters in real time. Researchers have been investigated for decades to try to find optimal combination to represent workloads accurately and then to predict the performance of the operators.

In this paper, the research was based on the assumption that the main electromyography (EMG) signals patterns for a particular task should be remain similar since it reflects natural requirements relating to the motor action associated with the task, and the secondary signal patterns associated with the main patterns should be evolving when the operator repeat the same task reflecting the operator's capabilities improvement. The content of our research includes the establishment of the experiment rig, designing and carrying out tasks, and analyzing the results. And first of all, the mapping between the state of hand motions and Surface Electromyography (sEMG) should be determined.

The EMG signal represents the electrical activity of muscles. EMG signals are usually detected via surface electrodes attached on the skin. EMG measures electrical currents that are generated in a muscle during its contraction. EMG signals can be used for a variety of applications including clinical applications, human-computer interaction and interactive computer gaming [3, 4]. Moreover, EMG can be used to sense isometric muscular activity which does not translate into movement. This makes it possible to classify subtle motionless gestures and to control interfaces without being noticed and without disrupting the surrounding environment [5].

2 Experiment

In order to obtain the signals regarding hand motion states, the sensors were placed on the right arm of the objects. The experiment about the sEMG acquisition of postures should be done to get the effects imposed by hand status. Taking into account the needs of experimental tasks, the directions of hand movements are divided into four categories, they are left, right, forward, and backward, respectively. The signal acquisition is separated into two phases: gesture acquisition and motion acquisition.

An sEMG gesture acquisition experiment should be carried out for the purpose of verifying the effect induced by the movement of arm on the sEMG signal characteristics during the posture retention of the arm muscles. And it is named static posture acquisition. Static posture acquisition is to explore the differentiation of the sEMG signal of particular muscles that maintain different postures of hand movements.

The experiment of motion data acquisition should be done to obtain the sEMG signals in the period of uniform motion caused by the right hand. The purpose of the test was to reveal the ability of a particular muscle EMG to recognize the periodic motion of the arm.

2.1 Participants

Ten volunteers (5 males and 5 females) were participated in this experiment with the age range from 22 to 26. They are all the postgraduates from School of Electronic Information and Electrical Engineering. Subjects have healthy body, normal eyesight or correction normal eyesight. Before the experiment, they were ensured to have adequate sleep without intense exercise. All of them were completely understood what would be done in the experiment and signed the consent form. The arm of the subject has no history of major injuries. Within 24 h before the experiment they did not take any irritating items.

2.2 Apparatus

Two kinds of devices were used throughout the experiment. The one is flying joystick (named Extreme 3D Pro) [6] attached Logitech trademark. The perfect ergonomic design with a custom twist-handle rudder relies its one-handed control resulting in a smaller device footprint. There are six programmable buttons on the base. Each programmable button can be configured to execute simple single commands or intricate macros involving multiple keystrokes, mouse events, and more. In this experiment, we only operate the rocker in the above four directions. (see Fig. 1)



Fig. 1. Myoelectric device (left) and flying joystick (right)

The other device is Delysis Trigno™ Wireless System [7], it was used to record the sEMG of the right forearm during operating the flying joystick. With a guaranteed transmission range of 40 m, a rechargeable battery lasting up to 8 h, and an intelligent sensor design, three-axis accelerometer is embedded in each EMG sensor. And the 64-channel synchronous signals can be outputted at the same time. Moreover, it can provide broader analysis data types. The high-frequency cutoff frequency of the amplifier is 5000 Hz, and the low-frequency cutoff frequency is 10 Hz. The range of signal amplitude of sEMG from 0mv to 10 mv, and its frequency ranges from 20–500 Hz [5]. The sEMG digitized by A/D converter at a sampling frequency of 2000 Hz in this experiment. The signals were amplified by a factor of 30 (see Fig. 1).

2.3 Muscle Selection and Electrode Placement

When selecting the muscle as the source of sEMG signals, the following four aspects should be taken into account [8]: the function of the muscle should be directly related

to the flexion and extension of the arm joint movement, the shape of the muscle should be relatively large enough, the muscle's position should be located in the shallow layer of their muscles, and the reliability of the acquisition results should be guaranteed when collecting the sEMG of the selected muscle.

According to the principle of local anatomy, the movement to bend and stretch the arm is mainly controlled by the upper arm muscle zone, holding fist and extensor are mainly controlled by the forearm muscle group. In this experiment, the flexor carpi ulnaris, flexor carpi radialis, brachioradialis, biceps brachii and palmaris longus are selected as the source of the sEMG (see Fig. 2).

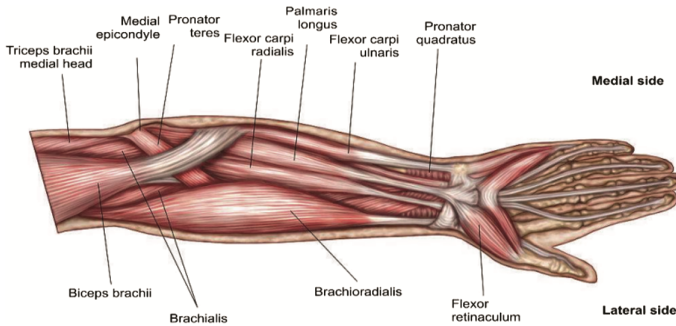


Fig. 2. Arm muscle anatomy

In addition, pay attention to the following points when placing the surface EMG electrodes: Surface electrodes should be placed in the place that is hardly affected by the crosstalk of adjacent muscle. For the majority of superficial muscles, the detection electrode should be placed in the middle of the tested muscle where is far enough from the other muscles. The direction of the arrow on the electrode should be parallel to the direction of the muscle spindles. Avoid placing the electrode on the tendon or near the tendon, avoid getting it close to muscle movement points and placing the electrode on the outer edge of the tested muscle is not recommended. To reduce the electrical impedance between the skin and the electrode, the forearm was shaved and cleaned with Ether and the electrodes were placed according to the anatomy showed on Fig. 2.

2.4 Procedure

Throughout the experiment, the subject sat on the ergonomic chair, keeping his/her upper body upright. Ergonomic seat is a kind of simplification of the cockpit seat, to maintain the basic functions. The angle of the backrest and the height of the seat are adjustable. In the experiment, the seat position is fixed and the subject kept the position and postures unchanged.

During the experiment of static posture acquisition, the subject grasped the joystick with his/her right hand, respectively, to the left, right, forward and backward, and recorded sEMG signals corresponds to four kinds of positions. The subjects were required to perform gestures in a sufficiently constant manner to eliminate the interference induced by the

difference of postures. Subjects would have some rest between any two actions to excluding the effect caused by muscle fatigue on sEMG signals.

As for the motion data acquisition. Subjects respectively completed the hand movements with exercise cycle of 2 s and 4 s during the experiment for five times. Moreover, the subject also need to complete the sEMG signal acquisition experiment with variable speed exercise in order to show the ability of the sEMG originated particular muscle to recognize the movement of the arm in a periodic jump, and in this case, the period of hand movement varied from 2 s to 4 s. Just as the process of static posture acquisition, muscle fatigue and subjects' postures should be taken into account.

2.5 Data Process

Signal Preprocessing. The sEMG signal is a complex physiological signal collected by placing the electrode on the muscle surface of the human body. The recorded electrical signal is preamplified and converted by A/D module and then sent to a signal processing module at the back end. The processed result reflects that the human body Muscle activity. The sEMG signal's energy is mainly concentrated in the following 1000 Hz [5]. From the Shannon sampling theorem, we can see that if we want to recover the sampled signal into the original signal and without distortion, the sampling frequency used should be greater than twice the maximum frequency of the original signal, so in this paper, 2000 Hz signal sampling rate was applied.

The sEMG signals show a train of motor unit action potentials corrupted with noise. Surface EMG signals, like most of the electrophysiological measurements, are frequently corrupted with three categories of noise [9], i.e. power line interference, white Gaussian noise, and motion artifact or baseline wandering. Noise contamination may compromise the efficacy of the EMG reduce the noise from surface EMG signals, among which the most simple and cost-efficient solution is to use conventional digital filters. For instance, the simplest method of removing narrow bandwidth interference from recorded signals is to use a linear recursive digital notch filter. Taking into account the efficiency of the algorithm and the actual needs of this article, the band-pass filter was used.

Activity Segment Detection. To explore the possibility of hand-motion encoding using multichannel EMG signals, signal-related segments of motion should be detected. During the execution of gesture actions, the sEMG signal detected by the sensors is called the active segment. Active segment can be straightforward to describe the sEMG signal for each action, it can be regarded as a gesture sEMG signal samples. To realize the recognition of gesture actions, it is necessary to detect the active segments of each gesture action from the continuous signals, that is, to determine the start and the end of each gesture action. The existing sEMG activity segment extraction algorithms include moving average method, short-time Fourier method, entropy theory method, etc. [10–12].

Considering the efficiency of the algorithm, in this paper, we use the moving average method based on threshold decision to detect the sEMG signals of the gesture actions. The moving average method uses a certain analysis window to calculate the timing signal, and applies the average of the window signal to represent this window signal, and the analysis window with time can predict the future signal direction. Active

segment detection judgment is made based on whether the energy of the sEMG signal sequence exceeds a preset threshold. Taking into account the different of the electrode positions caused by sEMG amplitude differences in different channels, active segment detection in all channels sEMG is based on the sum of the average. By selecting the appropriate parameters of the mean square value of each channel moving average, it is convenient to determinate the starting and ending points corresponding to the action. And the specific process is as follows:

- (a) Calculate the summation average of multichannel sEMG signals at time t , and then square the average signal to obtain the instantaneous energy sequence.

$$sEMG_{aver}(t) = \left[\frac{1}{C} \sum_{k=1}^C sEMG_k(t) \right]^2 \quad t < M \quad (1)$$

where C is the number of channels, M is the total number of signal sampling points.

- (b) Take the width of the active window $N = 100$ points (Equivalent to 50 ms signals length at 2000 Hz sample rate), deal with the squared signals movingly by averaging to get the value of moving average at point t .

$$sEMG_{MA}(t) = \frac{1}{N} \sum_{n=t}^{t-N+1} sEMG_{aver}(n) \quad t \leq M + N - 1 \quad (2)$$

- (c) Compare $sEMG_{MA}(t)$ and a definite threshold TH to determine the action signal. The signals with whose $sEMG_{MA}(t)$ is greater than TH and the length exceeds a certain set value are considered as signal segment, otherwise, the signals are known as noise. Described below using the formula:

$$sEMG_{rec}(t) = \begin{cases} sEMG_{MA}(t) & \text{if } sEMG_{MA}(t) \geq TH \\ 0 & \text{if } sEMG_{MA}(t) < TH \end{cases} \quad (3)$$

where $sEMG_{rec}(t)$ is the rectified signal at time t .

Sliding Window. After the raw sEMG signals were processed by band-pass filter, the sliding window was employed to deal with the signals. Explaining in detail, the signal of this section was represent by calculating root mean square (RMS) of the sampling points inside the window which was a certain length of the analysis window slid on the timing signal. The RMS value of the sEMG signal, as an index of the time domain of the EMG signal, represents the instantaneous electric power of the EMG signal and can represent the effective value of the muscle surface discharge. Modern research results show that the RMS waveform is similar to the linear envelope waveform of EMG signal and reflects the amplitude variation characteristics of sEMG signal in the time dimension. Its value is related to the synchronization of motion unit recruitment and excitement rhythm, and depends on the intrinsic relationship between the factors of the muscle load

and the physiological processes of the muscle itself. And it is often used to describe the state of muscle activity because of its good real-time performance [14]. Hence, we choose RMS as a parameter to evaluate the degree of dynamic muscle activity. By definition, the formula of RMS is as follows:

$$RMS = \sqrt{\frac{1}{T} \int_t^{t+T} sEMG^2(t) dt} \tag{4}$$

where $sEMG(t)$ is the sample value of the muscle surface signal at the time t , T is the length of time during a sampling period.

Determining the Range of the Threshold. After getting the activity section of signals, activity segments would be coding. That is Binarization. Binarizing the sEMG signals is actually finding the best step function to fit the signal curves. Here, the step function we use is as follows:

$$error = \sum_{i=1}^{N_1} \sum_{j=1}^C (Q_{ij} - R_{ij})^2 \tag{5}$$

where *error* is on behalf of the difference between the true value and the predicted value. N_1 is action mode, and its value is an integer and does not exceed the number of eight. Q_{ij} is the true value. R_{ij} is the predicted value. And R_{ij} can be described by the following formula:

$$R_{ij} = \begin{cases} 2 \times thre & \text{if } Q_{ij} \geq thre \\ 0 & \text{if } Q_{ij} < thre \end{cases} \tag{6}$$

where *thre* is the threshold of the binarization process.

3 Result and Discussion

In the signal processing software environment, this paper uses the band-pass filter to filter the signal. In the process of data acquisition of static postures, every subject were requested to keep the flying joystick in a constant manner toward four kinds of directions. Only two cases are shown here, the result of operation are respectively opened up Figs. 3 and 4. In the following pictures, there are two columns. And the left column is the raw sEMG signals after treating by band-pass filter. The five curves represent brachioradialis (abbreviated as Br), flexor carpi radialis (FCR), flexor carpi ulnaris (FCU), biceps brachii (BB) and palmaris longus (PL) from top to bottom, respectively. And the right column shows the root mean square corresponds to the left signals which is obtained by moving the sliding window.

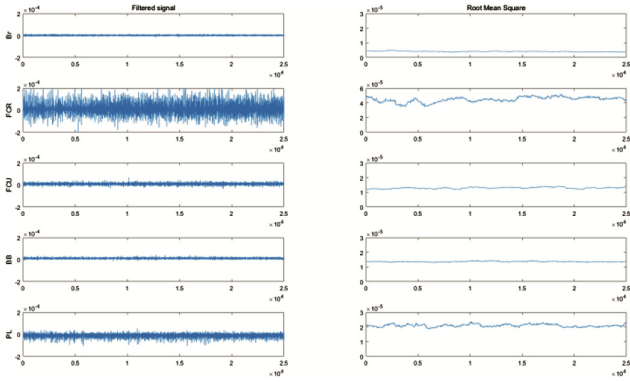


Fig. 3. The result of operation to keep joystick leftward

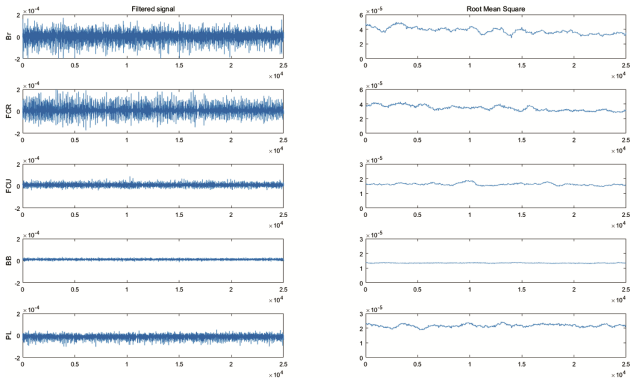


Fig. 4. The result of operation to keep joystick rightward

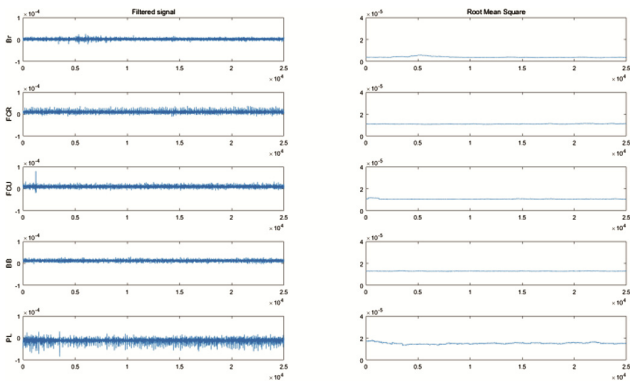


Fig. 5. The result of operation to keep joystick centre position

Combined with Figs. 3 and 4, we can see that five curves are generally stable with negligible minor fluctuations in the case of data acquisition of static postures. As the control group experiment, the Fig. 5 shown the result of keeping the right hand on the joystick without exertion. Observing the three pictures, the amplitude of the biceps brachii is almost unchanged. This result shows that the biceps brachii did not participate in exercise during the experiment. Comparing Fig. 3 with Fig. 4, the amplitude of palmaris longus is approximately equal. The voltage’s amplitudes produced by the other three kinds of muscles are different. Therefore, the sEMG signals is sensitive to changes in movement and the value of RMS can be employed as a kind of feature to identify different hand movements.

In the process of data acquisition of motion postures, the subjects manipulated the forearm to do periodic exercise. There were only shown the result of operation to the right and back in Figs. 6 and 7. From the two pictures, the true that the biceps brachii did not participate in motions was verification. And the signals show periodicity as the periodic exercise of the hand except for the signal induced by the flexor carpi radialis. The sEMG signals amplitude produced by the same muscles in different exercise modes were different. Hence, we can distinguish different actions by extracting different features. Furthermore, from the Fig. 8, we can know that the width of the action signals varies with the length of the action cycle, and the longer the action period, the wider the

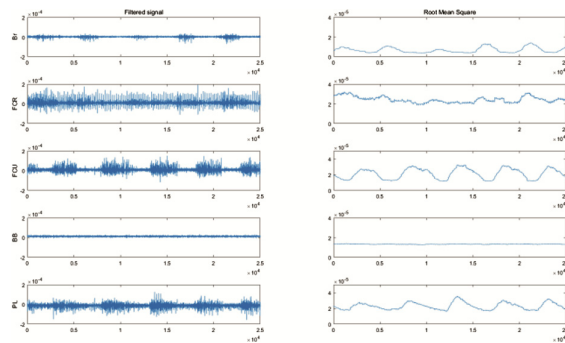


Fig. 6. The result of operation to keep periodic exercise backward

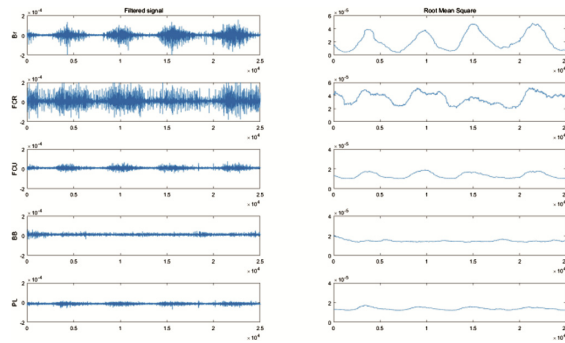


Fig. 7. The result of operation to keep periodic exercise right

envelope of the signal. In summary, sEMG signals obtained by carrying out the method mentioned above have strong ability and sensitivity of recognizing hand movements that get dynamically change.

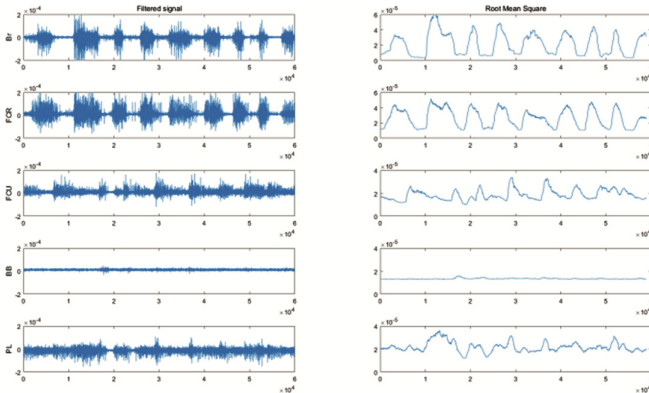


Fig. 8. The result of operation to keep periodic exercise changeable

In addition, the abscissa on the graph represents the number of points sampled, and the length between each two points corresponds to 0.5 ms on the time axis. The vertical axis of the graph is on behalf of sEMG signal amplitude, its unit of measurement is V.

For the activity section of signals, the sliding window was used to obtained the corresponding root mean square. Then we can get the error according to the Eq. (5) (Fig. 9). Here, Ten samples are randomly selected, and each sample includes 8 kinds of states that were described as above. From the Fig. 9, the range of optimal threshold could be obtained. And the value range is 1.2×10^{-5} – 1.7×10^{-5} .

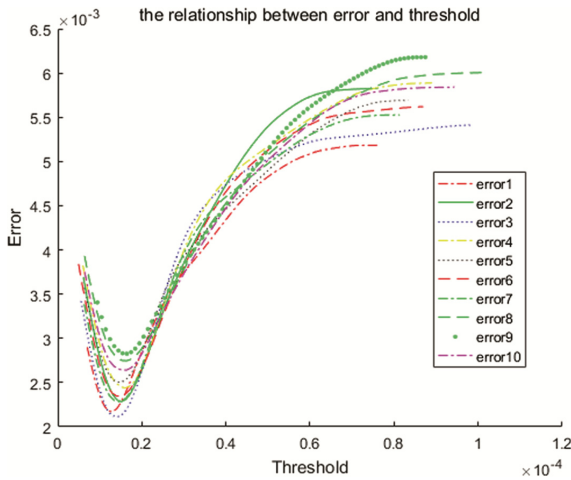


Fig. 9. The picture about error varied with the threshold

For further study the optimal threshold, the accuracy was calculated (shown in Fig. 10). The mode1, mode2, mode3 and mode4 are the state of hand motion in the case of dynamic gesture. And the mode5, mode6, mode7 and mode8 are the state of hand motion in the case of static gesture. The left picture is one result of coding, we call it Pattern one. The right picture is another result of coding called as Pattern two. From the Fig. 10, mode2 (downward) and mode3 (leftward) have the same coding and their coding are robust. As for the specific meaning of the pattern, we can know according to Table 1 as follow. When the value of threshold is close to the left of the abscissa, the accuracy of the Pattern One is higher. And when the value of threshold is close to the right of the abscissa, the accuracy of the Pattern Two is higher.

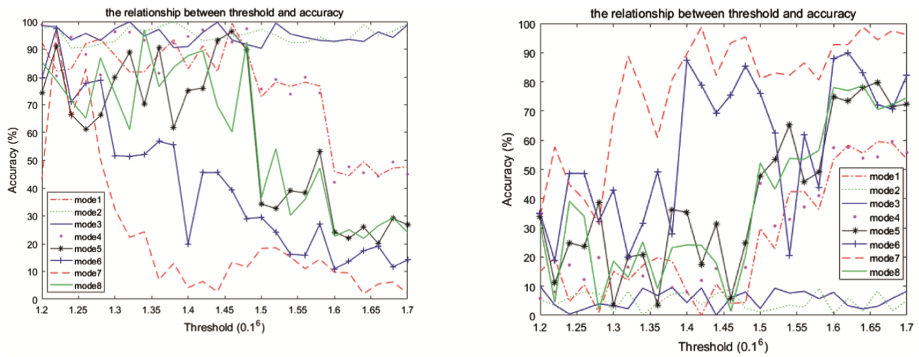


Fig. 10. The relationship between threshold and accuracy

Table 1. The meaning of the Pattern One and Two

| Pattern | Dynamic gesture | | | | Static gesture | | | |
|---------|-----------------|----------|----------|-----------|----------------|----------|----------|-----------|
| | Upward | Downward | Leftward | Rightward | Upward | Downward | Leftward | Rightward |
| One | 0111 | 0111 | 0111 | 1111 | 0111 | 1111 | 0111 | 1111 |
| Two | 0011 | 0111 | 0111 | 1100 | 0011 | 1101 | 0101 | 1101 |

Once the binarized threshold has been determined, we can get a series of encoding result according to Eq. (6), and the result of coding are partially displayed in Fig. 11.

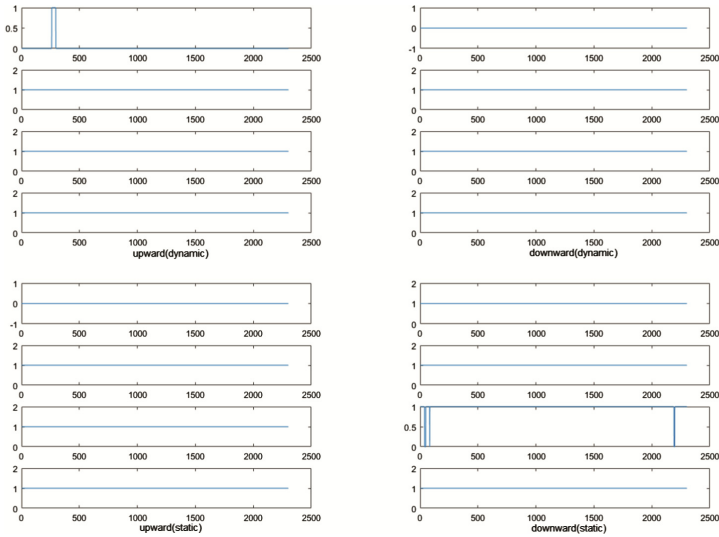


Fig. 11. The partial result of coding

4 Conclusion and Future Work

A raw sEMG signal contains more important information regarding the nervous system in useless form. In this paper, an experiment was designed to obtain the data of sEMG in different motion modes. The aim of this paper is to give detailed information about clearing up commonly associated noises and artifacts from sEMG signals, and to explore the relationship between the states of muscles of the forearm and sEMG signals. The result shows that the biceps brachii did not participate in exercise during the experiment and sEMG signals have strong ability and sensitivity of recognizing different hand movements. Further, the binarization method to code the sEMG signals was applied in this paper. The result shows that the choice of threshold has a great influence on the result of signal coding and we can not obtain proper coding only using the parameter of time domain. So, in the next work, suitable features including time domain, frequency domain and time-frequency domain features would be selected to better establish the correspondence for the specific and real experimental environment and platform mentioned above.

References

1. Zolghadri, A.: Early warning and prediction of flight parameter abnormalities for improved system safety assessment. *Reliab. Eng. Syst. Saf.* **76**(1), 19–27 (2002)
2. Hollands, J.G., Wickens, C.D.: *Engineering psychology and human performance*. Prentice Hall, New Jersey (1999)
3. Kim, J., Bee, N., Wagner, J., André, E.: Emote to win: affective Interactions with a computer game agent. In: *Lecture Notes in Informatics (LNI)*, vol. P-50, pp. 159–164 (2004)

4. Taneichi, T., Toda, M.: Fighting game skill evaluation method using surface EMG signal. In: 2012 IEEE 1st Global Conference, Tokyo, Japan, pp. 106–107. IEEE (2012). <https://doi.org/10.1109/gcce.2012.6379550>
5. Kim, J., Mastnik, S., André, E.: EMG-based hand gesture recognition for realtime biosignal interfacing. In: Proceedings of the 2008 International Conference on Intelligent User Interfaces, Gran Canaria, Canary Islands, Spain, pp. 30–38. DBLP (2008). <https://doi.org/10.1145/1378773.1378778>
6. <https://www.logitech.com/en-us/product/extreme-3d-pro-joystick>. Accessed 05 Feb 2018
7. <http://www.delsys.com/products/wireless-emg/>. Accessed 05 Feb 2018
8. Li, J.: Research on controlling methods of lower limb rehabilitation robot based on sEMG, pp. 24–31 (2013). [cnki.net](http://www.cnki.net)
9. Clancy, E.A., Morin, E.L., Merletti, R.: Sampling noise-reduction and amplitude estimation issues in surface electromyography. *J. Electromyogr. Kinesiol.* **12**(1), 1–16 (2002)
10. Fleischer, C., Wege, A., Kondakk, K., Hommel, G.: Application of EMG signals for controlling exoskeleton robots. *Biomed. Tech. (Berl)*. **51**(5–6), 314–319 (2006)
11. Daley, H., Englehart, K., Hargrove, L., Kuruganti, U.: High density electromyography data of normally limbed and transradial amputee subjects for multifunction prosthetic control. *J. Electromyogr. Kinesiol.* **22**(3), 478–484 (2012)
12. He, L.: Research on Human Machine Interface Based on EMG Signal. Southeast University, Nanjing (2006)
13. Hussain, M.S., Reaz, M.B.I., Mohd-Yasin, F., Ibrahimy, M.I.: Electromyography signal analysis using wavelet transform and higher order statistics to determine muscle contraction. *Expert Syst.* **26**(1), 35–48 (2009)
14. Jian, W.: sEMG signal analysis and its application research progress. **20**(4), 56–60 (2000)



Short Paper: Damage Mechanism and Risk Control on Kid's Sunglasses

Xia Liu¹(✉), Bisong Liu¹, Bao Liu², Youyu Xiao², and Yongnan Li²

¹ Quality Management Branch of China National Institute of Standardization, Beijing, People's Republic of China

liuxial010@163.com, liubs@cnis.gov.cn

² Nanjing Institute of Product Quality Inspection, Nanjing, People's Republic of China

gracyl00@163.com, gracyl010@163.com

Abstract. Kid's sunglasses have been the articles for children daily use. Therefore based on the children's physical and mental development characteristics and injury events in recent years, this paper analyzes the harm mechanism of nickel precipitation, transmission, high-temperature resistance and so on of the kid's sunglasses, parses the problems existed in standards and supervision of kid's sunglasses in China, and offers proposals for accelerating revision of national standards, strengthening consumption guide and enterprise supervision in order to improve the quality safety level of kid's sunglasses.

Keywords: Kid's sunglasses · Quality safety · Risk analysis and standards

1 Foreword

People regulate the light flux though the pupil size in the sun, but the eyes are hurt when the light intensity exceeds regulating capacity of the eyes. So the sunglasses are used in summer in some outdoor activities to relieve eye regulating fatigue or prevent injury from strong light stimulation. As for sunshading, the sunglasses, also called sun blinkers, is a vision care appliance to protect the eyes from strong sunlight stimulation. With the improvement of people's living and educational levels, the sunglasses also become the special ornament for beauty or personal style. Sunglasses can be divided into ordinary, polarized or specialized ones, etc. While the kid's sunglasses refer to the ones that are designed and manufactured specially for children's sunshading and ornament.

As a large country for sunglasses production, import and export, there are nearly 1000 sunglasses manufacturers in China currently, they are spread over Guangdong, Fujian, Zhejiang, Jiangsu, Shanghai and so on, their output is 40% of the total output in the world, the annual output value exceeds RMB8,000,000,000, and the annual volume of export is about USD600,000,000–700,000,000. At the same time, China also the consumption power of glasses, the consumer group of sunglasses is about 400,000,000 people.

2 Review of Injury Events Concerned Product Quality Safety

In recent years, reports about teenagers' pathopoiesis and injury caused by quality problems of kid's sunglasses are repeated. According to incomplete statistics, in 2009–2013, there are more than 40 quality safety events caused by the kid's sunglasses.

Case 1: In September 2011, a consumer feeds back that his 6 years old child has swollen, blisters and suppurating on his temples after wearing the sunglasses. In the hospital, it is confirmed that the reason is nickel in the spectacle frame.

Case 2: In June 2013, a 5 years old child named Yang Yang in Weihai, Shandong wears four kid's sunglasses in the shapes of Xiyangyang, Grandfather Sun, little frog and heart. After two weeks, the kid has photophobia, tears and gum in the left eye, the doctor diagnoses as keratitis. And the reason is Yang Yang often wears the inferior sunglasses.

Case 3: On June 12, 2013, a consumer in Meishan, Nanjing complaints that his 12 years old son's resin lens burst suddenly to hurt the eye, and several stitches are given in the hospital. The experts believe that the resin lens expand quickly when they go from low to high temperature suddenly, but the frame constrains them till they bust.

3 Main Hazards and Injury Mechanism

3.1 Characteristics of Kid's Visual Development

The 0–14 years old children are in the growth and development stage, including their eyes. The vision improves fastest in 3–5 years old, and the vision reaches the adult level at the age of 6; but before it, physiological hyperopia is the physiological character of children's eyes. For this reason, children under 6 shall not wear the sunglasses. In 7–10 years old, the children's eyeballs grows with age, their axis oculi extend gradually too, which reduces the degree of physiological hyperopia step by step. By the age of 7, the children's eyes approach the adult. And the children's eye size almost reaches the size of a adult at the age of 10. The eyeballs of the children in 10–15 years old still grow slowly, with the refractive status change constantly, which increases the risk of adverse effects caused by external factors.

3.2 Injury Mechanism

The main hazard factors to cause the above mentioned kid's sunglasses quality safety harmful events are the nickel release of the frame, the transmission property and the high-temperature resistance.

Nickel Release. As a heavy metal element, nickel may cause skin contact allergy. According to the medical evidences, contacting nickelic articles for long-term may cause skin allergy even carcinogenic. Extra nickel release from the kid's sunglasses frame may threaten health. For nickel ion may penetrate into the skin through the pore and sebaceous gland following sweat to cause allergy and inflammation of the skin, and the clinical manifestation is dermatitis and eczema. The clinical manifestation of the

nickel allergic dermatitis is pruritus, popular dermatitis or popular vesicular dermatitis with lichenification, even skin eruptions.

When the children wear the sunglasses in summer, the inner side of the glasses leg and the frame contact the skin for long-time, besides children are active and easy to sweat, so nickel element is absorbable through sweat (Fig. 1). In particular, the children's body apparatus are developing and vulnerable, so safety of wearing the sunglasses with a metal frame in high temperature cannot be ignored.

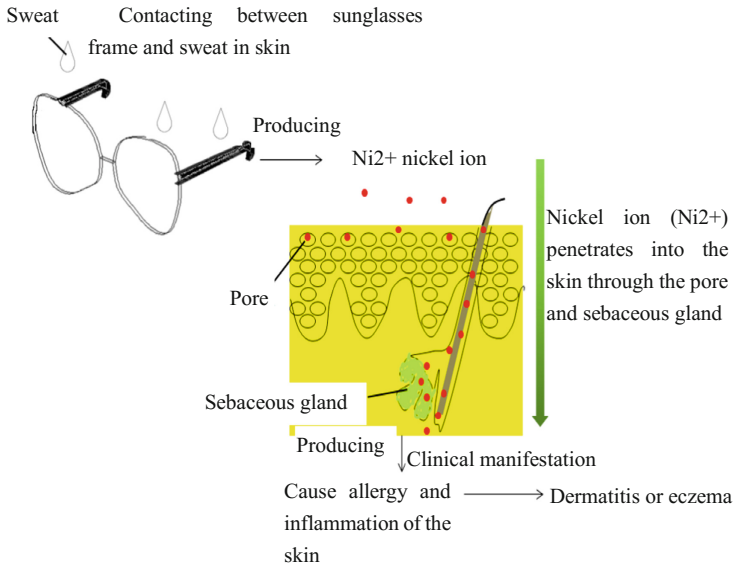


Fig. 1. Nickel precipitation and body absorption for a child wearing the sunglasses

Transmission Property. The transmission property is mainly expressed by the light transmissivity, including the visible light transmissivity and the ultraviolet spectrum transmissivity. Because children's vision is growing and developing, the kid's sunglasses with too low light transmissivity (Color too dark) can cover the children's eyes, restrain normal development of the children's vision and cause amblyopia. While a too

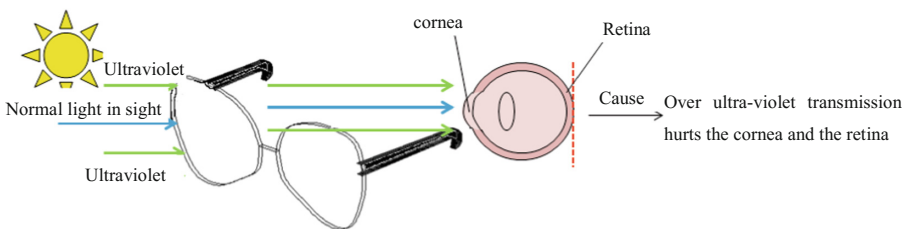


Fig. 2. Schematic diagram of hazard to children caused by the kid's sunglasses transmission properties

large ultraviolet transmission can hurt the eyes, especially the children's eyes are very sensitive to ultraviolet, so over ultraviolet transmissivity is easy to hurt the cornea and the retina (Fig. 2).

High-Temperature Resistance. The lens of the sunglasses are made from resin normally, in high temperature, the sunglasses resin lens expand by heating, during this procedure, their physical stability decreases largely, especially in the environment of excessive heat and cold, cracks or fractures are prone to producing, even burst, which may hurt the user's face and eyes (Fig. 3). Because children are active and playful, they often use the sunglasses as the toy. Many children wear the sunglasses for long time, which is very easily to occur lens fracture caused by a high temperature.

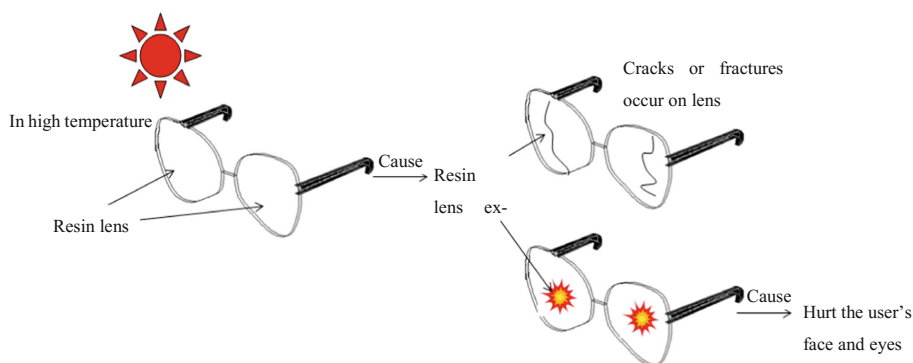


Fig. 3. Hazard caused by kid's sunglasses in high temperature

4 Standards Both in China and Abroad

4.1 Nickel Release

Ophthalmic optics – Spectacle frames – Requirements and test methods (ISO12870-2012) specifies that the limit of nickel release of the spectacle frames is $0.5 \mu\text{g}/\text{cm}^2/\text{W}$. And Ophthalmic optics – Spectacle frames – Method for the simulation of wear and detection of nickel release from coated metal and combination spectacle frames (ISO/TS 24348-2007) specifies that the nickel release from metal or alloy of frames directly contacting the user shall not be more than $0.5 \mu\text{g}/\text{cm}^2/\text{W}$. Eye and face protection – Sunglasses and related eyewear – Part 1: Sunglasses for general use (ISO12312-1, will be released soon) requires that design and manufacture of the sunglasses shall not endanger the users' health and safety, and the risk of users' skin injury caused by lens or frame material release shall be minimized. In addition, Korea Agency for Technology and Standards issues the amendment for safety and quality marks of industrial products, which adjusts the nickel release of the metal sunglasses and spectacle frames to less than $0.5 \mu\text{g}/\text{cm}^2/\text{W}$.

But standards concerned glasses in China currently do not consider the limit of the nickel release and the test method.

4.2 Transmission Property

Both ANSI Z80.3:1996 of USA and EN 1836:1997 of EU specify the transmission property of the sunglasses, and require the transmissivity of UVA shall not be more than 5%, and the transmissivity of UVB shall not be more than 1%.

China industrial standard Sunglasses (QB2457-1999) divides the sunglasses into 3 classes, the light colored sunglasses, the sun blinkers and special purpose sunglasses. Where the UVB transmissivity of the light colored sunglasses shall not be more than 30%, and the UVA transmissivity shall not be more than the transmissivity of visible light; the UVB transmissivity of the sun blinkers shall not be more than 5%, and the UVA transmissivity shall not be more than the transmissivity of visible light; while the UVB transmissivity of the special purpose sunglasses shall not be more than 1%, and the UVA transmissivity shall not be more than half of the transmissivity of visible light. As for the national compulsory standard, Spectacle lenses and related eye wear – Part 3: Transmittance specifications and test methods (GB10810.3-2006), the sunglasses are divided into 4 classes according to the transmittance, Class 1, 2, 3 and 4. Where the UVA transmissivity of Class 1, 2 and 3 shall not be more than 5%, the UVB transmissivity shall not be more than 1%; while the UVA transmissivity of Class 4 shall not be more than half of visible light transmittance, and the UVB transmissivity shall not be more than 1%. The kid's sunglasses in this paper refer in particular to the sun blinkers with the transmittances of Class 2 or 3.

4.3 High-Temperature Resistance

Currently, there is no relative requirements concerned the high-temperature resistance of the kid's sunglasses.

5 Suggestion and Solution

Accelerate Preparation and Revision of National Standards. Strengthen basic scientific research for relative standards concerned children products, and issue the standard about safety requirements for the kid's sunglasses in good time according to children's physical development characteristics. For example: Try to introduce the limit of nickel release requirement into the standard about the spectacle frames, especially the kid's sunglasses, Adornment—Provision for limit of baneful elements (GB28480-2012) can be referred or it can be quoted for nickel migration, thereby nickel release for the kid's sunglasses frames will be stricter than the adult products. At the same time, bring high-temperature resistance into the standard for the kid's sunglasses, specify the standard parameters or the detailed test methods strictly, in order to reduce the risk to children caused by products with bad high-temperature resistance.

Strength Consumption Guide. Strengthen propaganda of sunglasses usage, remind the consumers especially children's parents to prevent relative risk, select the sunglasses made by legitimate manufacturers, and take care of the children during using the product in order to avoid physical and mental damage to children.

Strengthen Supervision to Manufacturers. The relevant authorities shall strengthen supervision to manufacturers of the kid's sunglasses, ban all unlicensed illegal enterprises, enhance spot check for product quality, punish unqualified enterprises severely, increase cost of illegal business, and enforce the enterprises to improve product quality.

Acknowledgement. This paper has been funded by the national key research and development project "Research on key technical standards for quality and safety control of consumer goods" (2016YFF02022600), and the project "Research on common technology for integrative services by internet plus" (2017YFF0209604).

References

1. There is no relative standard in China though the kid's sunglasses are very popular on the market. *China Glass. Sci. Technol. Mag.* (09) (2015)
2. Ren, M.: Nearly 70% kid's sunglasses from online shopping are unqualified. *Qual. Explor.* (07) (2014)
3. Zhang, C., et. al.: Understanding international standard for sunglasses ISO 12312-1:2013. *China Glass. Sci. Technol. Mag.* (13) (2014)
4. TIANRUN RX LAB: Pioneer of one-stop customized service for sport sunglasses lens. *China Glass. Sci. Technol. Mag.* (11) (2015)
5. Xiao, L.: Strengthen market supervision and management to provide quality-assured glasses to consumers. *Ind. Measur.* (02) (1999)
6. Liu, H.: Model-based safety risk assessment method. *Comput. Eng.* (09) (2005)
7. Li, X.: Information security risk assessment model based on the danger theory. *J. Tsinghua Univ. (Sci. Technol.)* (10) (2011)
8. Zi, M.: Sunglasses, do you choose the right one. *Pop. Stand.* (8) (2012)
9. Anonymous. CPSC: Children's sunglasses recalled by axiom due to violation of lead paint standard. *M2 Presswire* (4) (2009)
10. Werner, J.S.: Children's sunglasses: caveat emptor. *Optom. Vis. Sci.* **68**(4), 318–320 (1991)



Affective Recognition Using EEG Signal in Human-Robot Interaction

Chen Qian, Tingting Hou, Yanyu Lu, and Shan Fu^(✉)

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

Abstract. Human-robot interaction is a crucial field in human factor field and mechanical arm operation is a widely used form in human-robot interaction. However, the mistaken operations caused by the affect inflution of operators are still one of the dominant reasons causing accidents. Because of the close link between affective state and human error, in this paper, we analyzed the EEG signal of five subjects operating mechanical arm and the track record of the mechanical arm movement. A combination label model including the subjective part and the objective part are proposed to reflect the real time affective state inflution. Additionally, in subsequent recognition experiment, the results indicate that the affect is a state of mind that requires a relatively longer period of time to be effectively represented and the frequency domain features are significantly more important than time domain features in affective recognition process using EEG signal.

Keywords: Affective recognition · Mechanical arm
Time domain features · Frequency domain features
Multi-scale sliding window

1 Introduction

Hitherto, mechanical arm, as a vital product in industry field, has been broadly used in medical, exploration, rescue field etc. However, the mistaken operation caused by human error, which should have been avoided, is still one of the dominant reasons causing accidents. As we all know, the performance of human has a close link to the cognitive state of human and to some degree affect is the main reason causing the change of the cognitive state. Therefore, one of the critical ways to avoid human error in mechanical arm operation is recognizing the affect during the operation process through detecting the physiological signal. Since R. W. Picard has defined *Affective Computing (AC)* [1] in 1995, affective computing has been a critical field in human-computer interaction area. There are numerous physiological signals which could reflect the cognitive state of human, such as Blood Volume Pressure (BVP), Skin Conductance Response (SCR), Respiration (RESP), Electrocardiogram (ECG), Electromyogram (EMG), Electro-corticogram (ECoG), Electroencephalogram (EEG), Heart Rate (HR), Oxygen

Saturation (SaO₂) and Surface Temperature (ST) [2]. In all physiological signals, EEG is no doubt the most capable signal directly reflecting the brain activity. Therefore this paper chooses to use 32 dry electrodes EEG signal acquisition equipment to obtain EEG signal data during mechanical arm operation.

There are lots of models which have been proposed to describe the affect, such as six basic emotions model proposed by Ekman et al. [3], eight basic emotions model proposed by Plutchik [4] and the valence-arousal scale proposed by Russell [5]. For simplifying this problem, this paper chooses to use the valence in the valence-arousal model of Russell to evaluate the emotion of the subjects. And the valence reflecting the positive or negative aspect of the subjects is enough to describe the cognitive state of the subjects.

Besides the affective model, the way to obtain the ground truth of the subjects cognitive state is also crucial. Almost all research in Affective Computing field use the self-assessment scores to estimate the true cognitive state of the subjects. However, even the subjects themselves could hardly to retell the exact affective state in the mechanical arm operation and using one single scores to estimate the cognitive state during the entire operation process is obviously not reasonable in detail. So this paper proposes to use objective and real time indexes to represent the cognitive state of the subjects. In this paper, we record the track of the end point of the mechanical arm and extract the features of the track to represent the fluctuation of the cognitive state of the subjects. Meanwhile, we assume that the workload and the time pressure could stimulate the affective change of the subjects, so we give a basic score, which reflects the affective state the subjects should be, and add the weighted track features scores to the basic score to reflect the fluctuation of the affective state.

In our experiment, we designed three levels of operating tasks in different difficulty to stimulate the affective state change. And the level of difficulty is determined by the workload and the time pressure. To eliminate the influence of unfamiliarity, one minute of free exploration is added before these three tasks and to eliminate the interaction of different tasks, a 30s reset time interval is added between the different tasks.

During the data process part, because of the low signal-to-noise ratio (SNR), the disturbance of EMG signal and the electromagnetic interference, the raw data have firstly been filter to the 1–64 Hz frequency band [6]. After normalization process, different scales sliding window are induced in extracting features. Because of the real-time label we obtain from the track mentioned above, it allows us to consider the data in single sliding window as one sample.

The feature extraction methods are detailed summarized in paper [6], we choose three time domain features and one frequency feature to represent the raw data according to the value of the weighted relative occurrence. And at the feature selection process, we apply Principal Component Analysis (PCA) to select the extracted features above. And at the classification design process, we apply Support Vector Machine (SVM), which is an effective classification discriminator, to predict the affective state.

Here is the reminder organization of this paper: Sect. 2 introduce the apparatus used in the experiment and the detail experiment protocol. Section 3

describes the data preprocess procedure, including EEG data preprocess and real time label data synthesis. Section 4 focuses on affective recognition methods, including multi-scale feature extraction, feature selection and classification design. Section 5 focus on the results of the experiment and the analysis of these results. Section 6 summarizes the conclusion of this paper.

2 Experiment Setup

2.1 Apparatus

There are 3 key equipments we use in our experiment: EEG Signal Acquisition Equipment, Mechanical Arm and Joystick. And there are 3 personal computers for communicating with EEG equipment, controlling movement of mechanical arm through joystick and showing the end point of the mechanical arm.

EEG Signal Acquisition Equipment: The EEG Signal Acquisition Equipment we apply is the Cognionics HD-72 Dry EEG Headset [7] (see in Fig. 1).



Fig. 1. Cognionics HD-72 Dry EEG Headset

Considering the difficulty of wearing EEG equipment and the comfort of the subjects, we choose 32 dry electrodes, according to the international 10–20 system, to obtain the EEG signal, which are showed in Fig. 2. The sampling rate is 500 Hz which is sufficient for obtaining EEG signal.

Mechanical Arm: The mechanical arm we apply is the Dobot Magician mechanical arm [8] (see in Fig. 3), because it supports reprogramming according to the need of users and the control precision (0.2 mm) is adequate for our experiment.

Joystick: The joystick we use is the flying joystick named Extreme 3D Pro [9] produced by Logitech (see in Fig. 4). The perfect ergonomic design with a custom twist-handle rudder relies its one-handed control resulting in a smaller device footprint. There are six programmable buttons on the base. Each programmable button can be configured to execute simple single commands or intricate macros involving multiple keystrokes, mouse events, and more. In our experiment, we only operate the rocker in six basic direction movement, which are left-right direction, front-behind direction and left-right rotation.

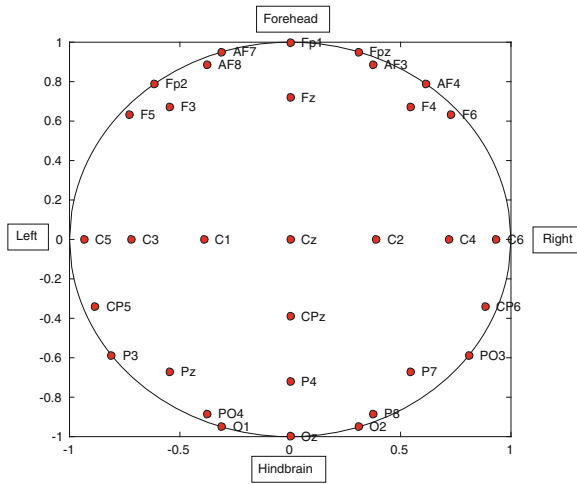


Fig. 2. 32 dry electrodes sensor location



Fig. 3. Dobot Magician mechanical arm



Fig. 4. Logitech Extreme 3D Pro joystick

In our experiment, we use Robot Operating System (ROS) to receive the control signal from joystick, release the control signal to mechanical arm, receive the position information of the end point of the mechanical arm and record the track information with real time mark point in a sampling rate of 10 Hz.¹ Meanwhile, real time EEG signal are recorded with mark information which could be used in subsequent time correcting process.

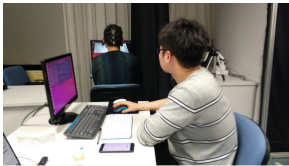
¹ Code url: <https://github.com/QCH1993/DobotMagician>.

2.2 Experiment Protocol

Five healthy participants (two females, three males), aged between 23 and 25, participated in the experiment. Before all experiments, all participants were asked to have adequate sleepness.

In our experiment, firstly, participants were asked to read the experiment procedures and notes and the experimenter would read out the procedures and notes for a second reminder. And the experimenter was also present there to answer any questions. At the beginning of each experiment, one minute of free exploration were given to remove the interference caused by unfamiliarity. And then the subjects were asked to operate the mechanical arm to touch the different color point on the desktop according to a certain order. We designed three levels of operating tasks in different difficulty (easy, medium and hard mode) according to the number and position of points and the time constraint. In easy mode, the subjects were asked to touch three colored points without time pressure. In medium mode, the subjects were asked to touch five colored points without time pressure. In hard mode, the subjects were asked to touch five colored points in 90 s. To eliminate the interplay between the three tasks, a 30 s break time for resetting was added after each task.

The experiment enviroment is showed in Fig. 5. The Fig. 5(a) is the overview of the entire experiment environment. The subject operation platform (see in Fig. 5(b)) is insulated from the mechanical arm platform and all the information helping the subjects to move the mechanical arm was from three camera set around the mechanical arm and on the end point of the mechanical arm (see in Fig. 5(c)). The screen interface presenting the information from cameras are showed in Fig. 5(d).



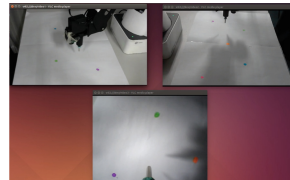
(a) overview



(b) the subject operating joystick



(c) the mechanical arm operation platform



(d) the screen showing camera information

Fig. 5. The experiment environment

3 Data Preprocess Methods

In data science field, data preprocess is the key procedure before any data analysis procedure. The quality of data preprocess directly affects the final accuracy of recognition. In EEG experiment, because of the low signal noise rate (SNR) and various interference, filtering and amplifying EEG signal is a crucial step before EEG data analysis. And in our experiment, we need to do some label data preprocess to obtain more objective real time label of EEG data because of the use of sliding window which would be mentioned in Sect. 4 and the need of combining self-assessment with objective data.

3.1 EEG Data Preprocess

For one subject experiment, the raw data are drawn in Fig.6(a). And then according to the track mark and EEG signal mark, the EEG signal and the

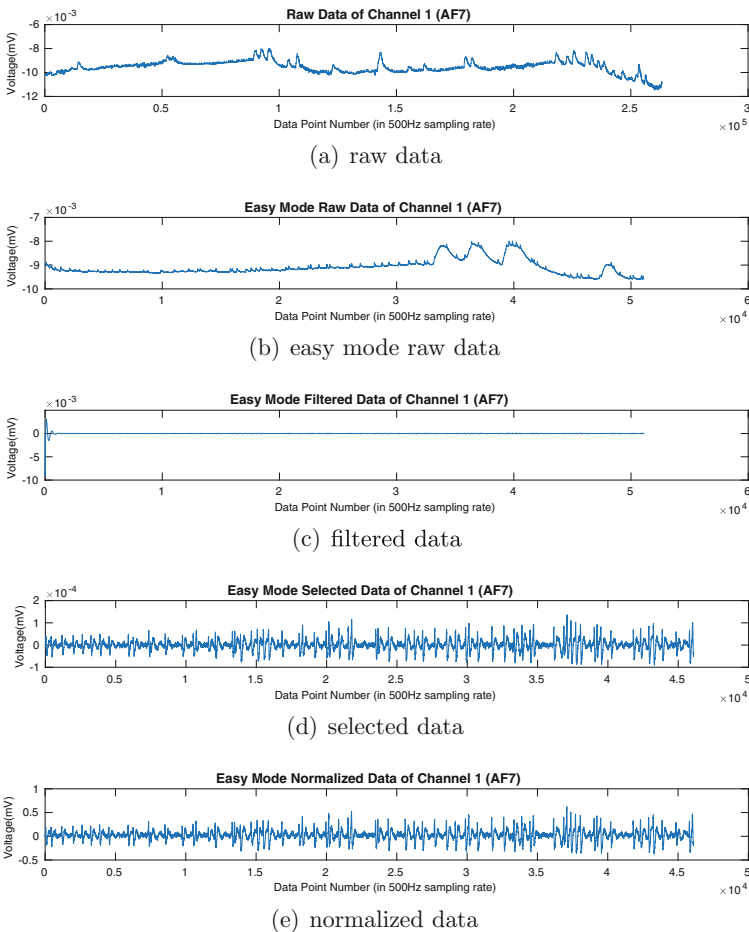


Fig. 6. Data preprocess wave charts

track were aligned in time. And we divided the entire EEG signal to easy mode, medium mode and hard mode procedure according to the mark information. The divided easy mode raw data, as an example, are showed in Fig. 6(b). Next, to remove the disturbance of EMG signal and the interference of various electromagnetic signal which are generally high frequency signals, we use band filter to filter out 1–64 Hz signal, which are the dominant frequency band of EEG signal (see in Fig. 6(c)). Because the EEG signal has extra low SNR, the filtered signal at the beginning and end part is unstable. Therefore we selected to remove the first and last 5 s data (see in Fig. 6(d)). And for the computing convenience, the selected data are normalized to $[-1,1]$ area according to all 32 channels signal by min-max normalization.

3.2 Real Time Label Data Synthesis

Label data includes three parts in total: the self-assessment of subjects, the difficulty level experimenters design and the performance the track reflects. The self-assessment is the subjective label data and the difficulty level and the performance are the objective data.

For the self-assessment, after the entire experiment were finished, each subjects were asked to fill a table about their affective state during operating different tasks. There were seven level the subjects could choose. “1” means the most positive affective state, “4” means neutral affective state and “7” means the most negative affective state. The greater the number, the worse the affective state. And then we used each score to weighted each task.

For the the difficulty level experimenters design, we simply gave the scores of difficulty according to the workload and the time pressure. In easy task, we asked the subjects to touch three points without time constraint, which means that we assumed the most positive affective state could be stimulated by this task. On the contrary, in hard task, we asked the subjects to touch five points with a ninety seconds constraint, which means that we assumed the most negative affective state could be stimulated by this task. We used “2.5”, “4” and “5.5” to represent the affective states that different difficulty levels could stimulate in our assumption.

For the performance, we recorded the tip trajectory of mechanical arm. As showed in Fig. 7, in easy mode, the subjects were asked to touch orange point, purple point and pink point in order from a random start point. We assumed that the subjects would feel positive when they moved the mechanical arm smoothly, so we counted the numbers of direction change, in selected sliding window mentioned in Sect. 4, as the estimation of the affective state of the subjects.

We use the formula below to calculate the final label.

$$L = \left[\frac{1}{2}L_{self-assessment} + \frac{1}{2}(L_{difficulty-level} + L_{direction-change}) + \frac{1}{2} \right] \quad (1)$$

where $[x]$ means the largest integer not exceeding x . And in this equation $L_{self-assessment}$ means the self-assessment score, $L_{difficulty-level}$ means the difficulty level score, and $L_{direction-change}$ determined by equation below:

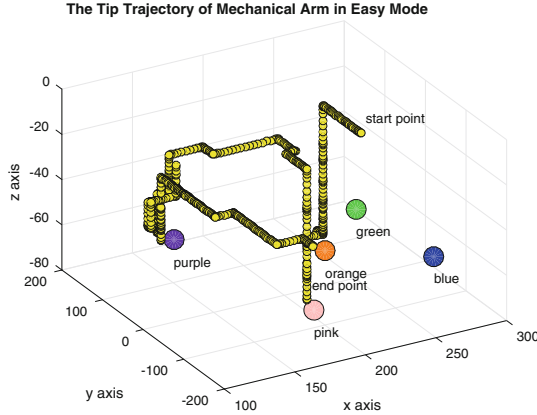


Fig. 7. The tip trajectory of mechanical arm in easy mode

$$L_{direction-change} = 3 \frac{N_{direction-change} - N_{min}}{N_{max} - N_{min}} - 1.5 \quad (2)$$

where $L_{direction-change}$ means the direction change score, among them, $N_{direction-change}$ means the number of direction changes, N_{min} means the minimum of $N_{direction-change}$ and N_{max} means the maximum of $N_{direction-change}$.

4 Affective Recognition Methods

Traditional pattern recognition process includes 4 steps: data preprocess, feature extraction, feature selection and classification design. In this paper, we use several classic methods in feature extraction, feature selection and classification design procedure to explore their performance in our experiment data.

4.1 Multi-scale Feature Extraction

The multi-scale feature extraction procedure are divided in two steps: (1) sliding window selection and (2) feature extraction.

Sliding Window Selection

Because uncertain relation between time interval and affective state change, we chose every 2s between 4s to 30s as the length of sliding window and 0.2s as the length of stride to extract multi-scale features.

Feature Extraction

According to the paper [6], we chose 18 dimensions time domain features, which are statistical features, higher order crossings (HOC) features and fractal dimension feature, and 53 dimensions frequency domain features, which are band power, bin power and the ratio of mean band powers β/α .

Time Domain Features

In this paper, we selected 3 classic time domain features in recognition process: statistical feature, higher order crossing (HOC) features and fractal dimension (FD) feature.

There are seven *statistical features* representing the raw EEG time series. These are:

- (a) Mean: $\mu_\eta = \frac{1}{T} \sum_{t=1}^T \eta(t)$
- (b) Power: $P_\eta = \frac{1}{T} \sum_{t=-\infty}^{\infty} |\eta(t)|^2$
- (c) Standard deviation: $\sigma_\eta = \sqrt{\frac{1}{T} \sum_{t=1}^{T-1} (\eta(t) - \mu_\eta)^2}$
- (d) 1st difference: $\delta_\eta = \frac{1}{T-1} \sum_{t=1}^{T-1} |\eta(t+1) - \eta(t)|$
- (e) Normalized 1st difference: $\bar{\delta}_\eta = \frac{\delta_\eta}{\sigma_\eta}$
- (f) 2nd difference: $\gamma_\eta = \frac{1}{T-2} \sum_{t=1}^{T-2} |\eta(t+2) - \eta(t)|$
- (g) Normalized 2nd difference: $\bar{\gamma}_\eta = \frac{\gamma_\eta}{\sigma_\eta}$

To find more robust features and pattern of EEG, Petrantonakis and Hadjileontiadis proposed *higher order crossings (HOC)* features [10]. Therefore we applied HOC to represent the raw EEG time series. The higher order series could be calculated by the formula below:

$$H_k\{\eta(t)\} = \nabla^{k-1}\eta(t) \tag{3}$$

where ∇ means $\eta(t) - \eta(t-1)$. Therefore, the features of HOC could be calculated by counting the sign changes or equation below:

$$F_k = \sum_{t=1}^{T-k} \psi(H_k\{\eta(t)\}H_k\{\eta(t+1)\}), k = 1, 2...10 \tag{4}$$

where $\psi(x)$ is a section function which is defined below:

$$\psi(x) = \begin{cases} 0 & \text{if } x \geq 0 \\ 1 & \text{if } x < 0 \end{cases} \tag{5}$$

The *fractal dimation (FD)* as a feature measuring the complexity is widely used. There are many methods computing the FD feature. In this paper, we chose to use the Higuchi algorithm [11] to caculate the FD feature. To compute the FD feature, the EEG series is rewritten as:

$$\{\eta(p), \eta(p+q), \eta(p+2q), \dots, \eta(q + [\frac{T-q}{q}]q)\}, p = 1, 2, \dots, q \tag{6}$$

where $[x]$ means the largest integer less than x. Then we could define the series $M_p(q)$ as below:

$$M_p(q) = \frac{T-1}{[\frac{T-p}{q}]q^2} \sum_{k=1}^{[\frac{T-p}{q}]} |\eta(p+kq) - \eta(p+(k-1)q)| \tag{7}$$

Therefore we could further define the average:

$$F(q) = \frac{1}{N_p} \sum_{p=1}^{N_p} M_p(q) \quad (8)$$

where N_p means the maximum of p . According to paper [11], we knew that $F(p)$ is proportional to $p^{-F_{FD}}$ (where F_{FD} means the feature of FD). Therefore we could assumed:

$$F(q) = r \cdot p^{-F_{FD}} \quad (9)$$

when we log the equation, we could obtain:

$$\log(F(q)) = \log(r) - F_{FD} \log(p) \quad (10)$$

Therefore we could obtain FD feature F_{FD} by the slope of $\log(F(q))$ to $\log(p)$.

Frequency Domain Features

For time series signal, spectrum analysis is an important method extracting features. Through short-time fourier transform (STFT) [12], which is a more robust method, we could get the power value as a function of time and frequency: $p(t, f)$. Further, the relation of the power and the frequency could be written as below:

$$P(f) = \sqrt{\sum_{t=0}^T [p(t, f)]^2} \quad (11)$$

As we know, effective EEG signals exists in low frequency band. Further, this valid low frequency band could be divided to five smaller parts: δ (1–4 Hz) band, θ (4–8 Hz) band, α (8–12 Hz) band, β (12–30 Hz) band, γ (30–64 Hz) band. Therefore, for each band on $[f_a, f_b]$, *STFT band power* could be calculated by the formula below. Additionally, the ratio of mean band powers $\frac{\beta}{\alpha}$ was computed.

- (a) Mean: $\mu_P = \frac{1}{f_b - f_a} \sum_{f=f_a}^{f_b} P(f)$
- (b) Minium: $P_{max} = \max(P(f)), f \in [f_a, f_b]$
- (c) Maxium: $P_{min} = \min(P(f)), f \in [f_a, f_b]$
- (d) Variance: $var(P) = \frac{1}{f_b - f_a} \sum_{f=f_a}^{f_b} |P(f) - \mu_P|^2$

Similarly, we could divide the frequency band in a higher resolution to obtain *STFT bin power* features. We divided the 1–64 Hz band into 32 subband, which means 32 ($\Delta f = 2 \text{ Hz}$) bands are extracted for further processing. And then, for each Δf band, we could caculate the STFT bin power feature by the formula above.

As showed in Fig. 8, We concatenate the time domain features and the frequency domain features as the representation of original raw data. Therefore, the features number of one channel of one sample is 71.



Fig. 8. Feature vector composition

4.2 Feature Selection

Principal component analysis (PCA), proposed by Pearson [13], is a effective and classic method to reduce dimensions of original matrix through keeping the lower order principal components. Therefore we use PCA to select the original features.

After the feature extraction steps, we could concatenate 32 channels feature to a long features vector and use all samples to constitute a feature matrix $\mathbf{M} = (\phi_1, \phi_2, \dots, \phi_N)$, where ϕ_i means one sample features vector, N means the number of samples.

Then we use the samples to estimate the mean and the covariance matrix:

$$\bar{\phi} = \frac{1}{N} \sum_{i=1}^N \phi_i \tag{12}$$

$$\Sigma = \frac{1}{N} \sum_{i=1}^N (\phi_i - \bar{\phi})(\phi_i - \bar{\phi})^T = \frac{1}{N} \mathbf{X} \mathbf{X}^T \tag{13}$$

Then we could caculate the eigenvectors (μ_i) and the eigenvalue (λ_i) of Σ . The value λ_i means the importance of the corresponding eigenvector μ_i . We sorted these eigenvalues in descending order to obtain the descending eigenvalue series: $\lambda'_1, \lambda'_2, \dots, \lambda'_i, \dots, \lambda'_L$, where L is the number of features. Correspondingly, the eigenvectors could be rearranged: $\mu'_1, \mu'_2, \dots, \mu'_i, \dots, \mu'_L$. Then we define the information integrity I by the formula below:

$$I = \frac{\sum_{i=1}^k \lambda'_i}{\sum_{i=1}^L \lambda'_i} \tag{14}$$

In this paper, we chose

$$K = \arg \max_k I(k) > 0.99 \tag{15}$$

Then K is the feature dimension of feature matrix after dimensionality reduction. And the mapping matrix is $M_{PCA} = (\mu'_1, \mu'_2, \dots, \mu'_K)$.

4.3 Classification

Support vector machine (SVM) is a effective classification proposed by Cortes and Vapnik [14]. Because the advantages of SVM in non-linear and high-dimensional pattern recognition, we applied it in our experiment as the classification.

When we classify data, each sample data is composed of a feature vector and a label value: $D_i = (\mathbf{x}_i, y_i)$, where \mathbf{x}_i is the feature vector in high dimension, y_i is the category the sample belongs. In particular, the classification problem of two category is taken as an example, and the multi-classification problem could be solved as the combination of multiple two classification problems. Specially, in two classification problem, we define the distance between one sample and a hyperplane is η_i , and $\eta_i = y_i(\boldsymbol{\omega}\mathbf{x}_i + \mathbf{b})$. And to normalize the distance, $\boldsymbol{\omega}$ and \mathbf{b} are replaced by $\frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|}$ and $\frac{\mathbf{b}}{\|\boldsymbol{\omega}\|}$. The normalized distance could be rewritten as

$$\delta_i = \frac{|g(\mathbf{x}_i)|}{\|\boldsymbol{\omega}\|} \tag{16}$$

where $g(\mathbf{x}_i) = \boldsymbol{\omega}\mathbf{x}_i + \mathbf{b}$. Because the relation between δ and misclassification number N is

$$N \leq \left(\frac{2R}{\delta}\right)^2 \tag{17}$$

where $R = \max\|\mathbf{x}_i\|, i = 1, 2, \dots, n$, the larger the distance δ is, the smaller the N is. Therefore, the optimizing equation is

$$\begin{aligned} \max \quad & \delta \\ \text{s.t.} \quad & y_i(\boldsymbol{\omega}\mathbf{x}_i + \mathbf{b}) - 1 \geq 0, i = 1, 2, \dots, n \end{aligned} \tag{18}$$

Further, according to Eq. 16, the optimizing target could be rewritten as below:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{\omega}\|^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{\omega}\mathbf{x}_i + \mathbf{b}) - 1 \geq 0, i = 1, 2, \dots, n \end{aligned} \tag{19}$$

Because $\boldsymbol{\omega}$ determined by sample data, $\boldsymbol{\omega}$ could be assumed as:

$$\boldsymbol{\omega} = \sum_{i=1}^n a_i y_i \mathbf{x}_i^T \tag{20}$$

a_i would be unequal to zero only when the sample is on the closest hyperplanes. In other word, hyperplanes are supported by the sample points which are close to these hyperplanes. Therefore, discriminant function could be written as:

$$g(\mathbf{x}) = \boldsymbol{\omega}\mathbf{x} + \mathbf{b} = \sum_{i=1}^n a_i y_i \mathbf{x}_i^T \mathbf{x} + \mathbf{b} \tag{21}$$

Since the samples on hyperplanes are known, \mathbf{b} could be caculated by $y_i(\boldsymbol{\omega}\mathbf{x}_i + \mathbf{b}) - 1 = 0$ When test sample \mathbf{x}_{test} need to be classify, the value of $g(\mathbf{x}_{test})$ decides the category \mathbf{x}_{test} belongs.

5 Results and Discussion

In our recognition experiment, we applied the ten-folder cross-validation method to test the accuracy robustness of the recognition algorithm. We randomly selected 10% original data as the test data, and chose the rest as the train data, and this process has been done 10 times for each specific parameter pair to eliminate the accidental errors.

5.1 Real Time Label Data Synthesis

As showed in Fig. 9, an example of the self-assessment score, the difficulty level score and the performance score was depicted. The trend of the difficulty level score and the self-assessment is basically the same.

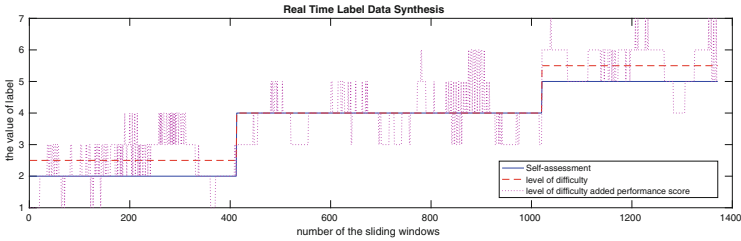


Fig. 9. Real time label data synthesis

After adding the performance score, as we assumed that the affective state is unstable even in the same task, there are some real time inflection which meets the need for obtaining real time label data to study the influence for recognition accuracy when the multi-scale sliding window were used.

5.2 Multi-scale Sliding Window

In our recognition experiment, multi-scale sliding windows were applied into extracting features. Because the detecting window of the performance feature extraction in real time label data synthesis is 2 s long, the scale of the sliding window should be longer than 2 s. Therefore, we chose every 2 s time interval from 4 s to 30 s to explore the relation between the accuracy and the length of the sliding window.

As showed in Fig. 10, the recognition accuracy tends to increase slightly as a whole when the sliding window grows longer. Specifically, when the length of the sliding window is 4 s, the recognition accuracy using frequency domain features is about 65%, which is obviously lower than the other length of sliding window. And when the length of selected window is longer than 20 s, the recognition accuracy using frequency domain features is almost stable at around 80%. This phenomena indicates that the affective state is a more stable state of mind that requires a longer period of time data to represent.

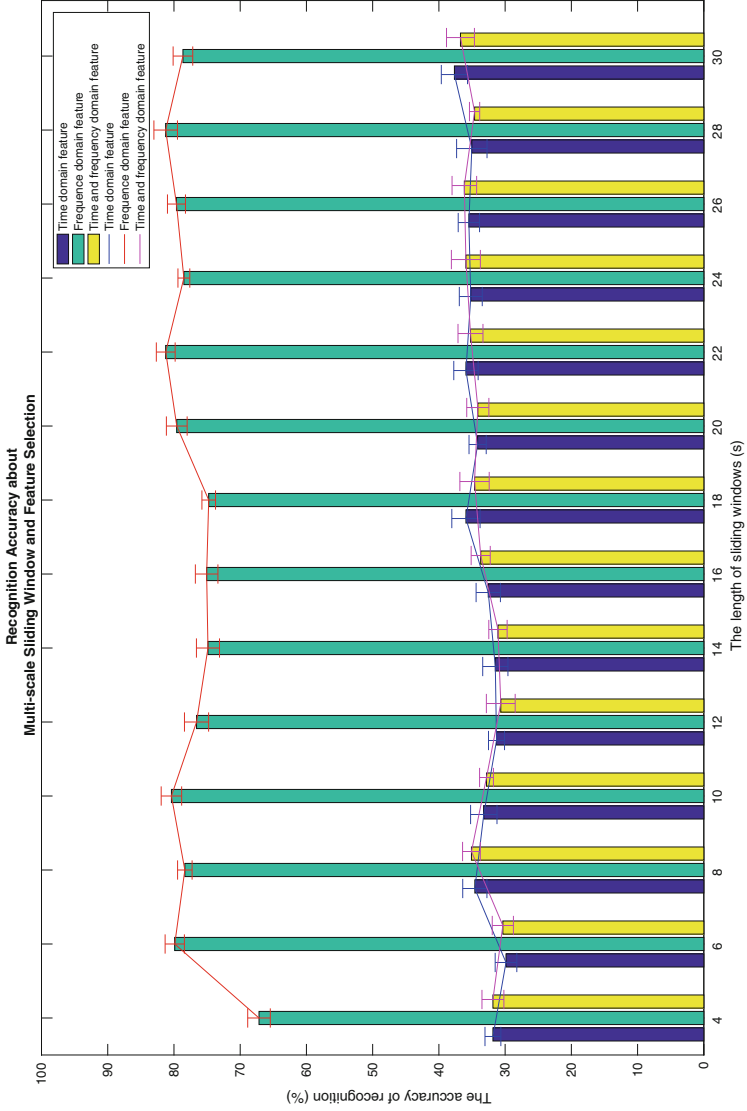


Fig. 10. Recognition accuracy about multi-scale sliding window and feature selection

5.3 Feature Selection

As showed in Fig. 10, when we fix the parameter of the sliding window length, the accuracy is highest in using frequency domain feature. And the accuracy is about the same in using time domain feature and the combination of time and frequency feature. This phenomena indicates that PCA could not select the more important feature componets beacuse PCA reduces the dimension of feature map just by using the information of covariance. Therefore, the PCA which could barely use the label information is unsuitable for EEG signal process. And the higher recognition accuracy also indicates that frequency domain features are more important for affective recognition by using EEG signal.

6 Conclusion

In this paper, we combine the objective evaluation to the self-assessment to obtain real time label of the affective state. The result of combination reflects the influction of affective state. And the recognition experiment indicates that the affect is a state of mind that requires a longer period of time to be effectively characterized and recognized. In subsequent affective recognition experiments, the results shows that relatively longer EEG data is more appropriate for affective recognition. Meanwhile, the recognition experiment also illustrates that the frequency domain features are significantly more important than the time domain features. In future EEG analysis work, relatively long frequency domain features might be a more preferred option.

References

1. Picard, R.W.: *Affective Computing*. MIT Press, Cambridge (1997)
2. Khosrowabadi, R., Wahab, A., Ang, K.K., et al.: Affective computation on EEG correlates of emotion from musical and vocal stimuli. In: 2009 International Joint Conference on Neural Networks, IJCNN 2009, pp. 1590–1594. IEEE (2009)
3. Ekman, P., Friesen, W.V., O’sullivan, M., et al.: Universals and cultural differences in the judgments of facial expressions of emotion. *J. Personal. Soc. Psychol.* **53**(4), 712 (1987)
4. Plutchik, R.: Emotions: a general psychoevolutionary theory. *Approaches Emot.* **1984**, 197–219 (1984)
5. Russell, J.A.: A circumplex model of affect. *J. Personal. Soc. Psychol.* **39**(6), 1161 (1980)
6. Jenke, R., Peer, A., Buss, M.: Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* **5**(3), 327–339 (2014)
7. Documents of the Cognionics HD-72 Dry EEG. <http://www.cognionics.com/images/docs/HD72.pdf>
8. Specification of the Dobot Magician mechanical arm. <https://www.dobot.cc/dobot-magician/specification.html>
9. Features of The Extreme 3D Pro joystick. <https://www.logitechg.com/en-us/product/extreme-3d-pro-joystick#featuresAnchor>

10. Petrantonakis, P.C., Hadjileontiadis, L.J.: Emotion recognition from EEG using higher order crossings. *IEEE Trans. Inf. Technol. Biomed.* **14**(2), 186–197 (2010)
11. Liu, Y., Sourina, O.: Real-time fractal-based valence level recognition from EEG. In: Gavrilova, M.L., Tan, C.J.K., Kuijper, A. (eds.) *Transactions on Computational Science XVIII. LNCS*, vol. 7848, pp. 101–120. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38803-3_6
12. Lin, Y.P., Wang, C.H., Jung, T.P., et al.: EEG-based emotion recognition in music listening. *IEEE Trans. Biomed. Eng.* **57**(7), 1798–1806 (2010)
13. Pearson, K.: LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2**(11), 559–572 (1901)
14. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995)



Research on Test of Anti-G Suits Airbag Pressure

Ding Yi¹, Zhaowei Zhu², Wang Yandong¹,
Zhang Zhongji¹, Song Kaiyuan¹, and Ding Li²(✉)

¹ Experimental School of Beihang University, Beijing, China

² School of Biomedical Engineering, Beihang University, Beijing, China
ding1971316@buaa.edu.cn

Abstract. In order to solve the problem of overloading, the common international practice is to use anti-load equipment (anti-load suits). This article focuses on the role of capsule anti-Dutch suit pressure on pilots. Considering the actual situation of overload, if the absolute value of overload is too large and the applied pressure is too small, it will lead to imperfect overload and cause flight safety hazard. Conversely, if the absolute value of overload is too small and the pressure exerted on it will lead to blood circulation disorder Serious or even tissue necrosis occurs. Therefore, it is of great value to study the numerical relationship between cystic pressure and body pressure. In this paper, the relationship between capsule anti-static clothing pressure and body pressure is explored by experimental method. Wearable body surface pressure measurement device is designed based on the membrane pressure sensor. The research shows that: (1) there is a correlation between capsule pressure and body pressure, the absolute value of the body pressure can be estimated by the capsule pressure; (2) The corresponding relationship between the capsule pressure and the body pressure of different parts is different, The corresponding relationship between gauge pressure is also a certain difference.

Keywords: Anti-G suits · Pressure · Pilot · Pressure test experiment

1 Introduction

With the development of modern aviation, the flight capability of high-performance fighter jets has been greatly improved [1]. Overload, fast growth of overload and long overload time have become their important features. Pilots will experience significant acceleration during flight and in the event of overload, centrifugal force applied from the head to the foot forces the blood to the lower body [2, 3]. If the pilot's muscular structure does not adjust well, There will be gray-shaded, black-and-visual issues of great visual impact on the flight [1]. In order to deal with this kind of problem, scientists invented anti-Dutch pilot clothing as early as World War II, and later evolved into a compensatory service [4]. The principle of compensatory service worn by pilots was to pressurize the human lower extremities and abdomen through clothing, To maintain the effective circulation of the body's head blood volume and reduce the adverse effects of

overload on the human cardiovascular system in order to achieve its anti-Dutch effect [5, 6]. However, there are still many problems with pilots compensating clothes: (1) many connections, complex structures and troubles in testing; (2) susceptible to wearing apparel, and poor real-time data testing [7, 8]; (3) There exists accuracy problem in testing. Therefore, on the basis of the previous studies, we improved the body surface pressure test of pilot compensation service [9]. We use RFP membrane pressure sensor, through the establishment of the test circuit system to obtain the resistance of the sensor changes, so as to obtain in the actual test wearing a compensatory pilot wearing the body surface pressure size. The wearable garment stress testing equipment we developed not only improves accuracy, but also makes testing easier than ever before.

2 Methods

In this study, RFP membrane pressure sensor was used to five parts (left and right thighs, left and right lower leg, abdomen) balloon pressure measurement and calibration to get more accurate pressure and pressure changes in the data, and thus for the future design of the pilot's clothing to do Make a little contribution.

2.1 Measuring Principle

RFP film pressure sensor consists of two thin polyester film, the inner surface of the two films contains conductors and semiconductors. The basic principle of the test is that by applying pressure to the surface of the RFP film, the resistance of the semiconductor decreases as the pressure increases, so different pressures will correspond to different resistances. In this way, by establishing a circuit sensor resistance system, you can know RFP membrane pressure sensor suffered the size of the pressure. When the sensor is used to test the pilot's 5 positions by the balloon pressure value, as long as you know the resistance of the pressure sensor, you can get different parts of the pilot suffered pressure and pressure changes.

2.2 Measurement Methods

- a. The pressure sensor has been sewn five stress test strap tied to the thigh (left and right), leg (left and right) and abdomen, and connected with multi-channel pressure obtaining instrument. At the same time the pressure acquisition instrument is access to the computer, and open the appropriate software on the computer, record the static pressure data.
- b. Increasing the access to gas, continuous recording of dynamic pressure data, but also record changes in the input pressure value.
- c. The pressure data collected by the system should correspond with pressure values recorded respectively.
- d. Experimental measurement device is shown in Fig. 1.



Fig. 1. Measuring equipment

2.3 Calibration Methods and Data Processing

Calibration Method

Calibrate the sensor with a press for every 50 g of boost (Figs. 2 and 3). Set the data measured by the sensor to y and press to x , and establish the linear relationship between x and y . Design programming experiment software According to its linear relationship (piezoelectric sensors have been calibrated by the manufacturer and provide the calibration data).



Fig. 2. Sensor calibration press

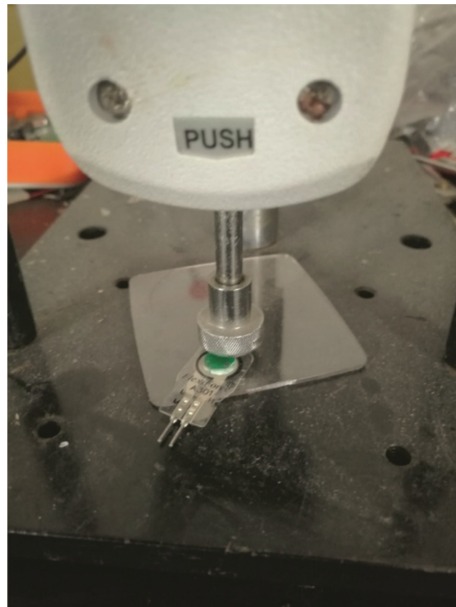


Fig. 3. Press calibration RFP film sensor

Data Processing

The data obtained from continuous testing is stored in Excel format with good notes. Calculate the slope and error when the pressure is rising by using the Excel to calculate the average of the fifteen measured data from the five sites. Assuming that the change of the data obtained by the pressure tester is y , the variation of the pressure unit of the gas introduced is x , and the linear relationship between y and x is established by using Excel to obtain the relationship between the actual compensated pressure and the body surface pressure. To prevent pilots in the high-speed overload dizziness situation to provide effective help.

2.4 Significance

Through the improvement of the sensor, to achieve more accurate and more convenient compensation service pressure measurement. The improved sensor can directly touch the skin, reducing the impact of dress on the experiment; RFP film pressure sensor consists of two thin polyester film, the inner surface of the two films laying conductor and semiconductor. Compensatory service Internal piezoelectric sensors on both sides of the layer to add a thin layer of elastomers, effectively help to absorb the error introduced by the force distribution, improve test accuracy; the use of multi-channel pressure sensors, each sensor corresponds to a channel, the data more accurate, more operational Convenience.

3 Results and Discussion

3.1 Detailed Description of the Research Process

In this study, we tested and recorded a total of 4 pilots pressure values at different press with compensated clothes, and also recorded the actual value of the gas pressure. The data collected by the pressure harvester is compared with the air pressure of the actual gas, and the relationship between them is found out.

3.2 Research Findings and Conclusions

This study mainly focuses on two points: (1) there is a correlation between capsule pressure and body pressure, the absolute value of the body pressure can be estimated by the capsule pressure, the effect of different body pressures on the blood circulation is different So as to dynamically update the capsule pressure according to the need of the flight so as to ensure the life safety of the pilot in the overload state; (2) The corresponding relationship between the capsule pressure and the body pressure of different parts is different, The corresponding relationship between gauge pressure is also a certain difference, so pilots should be obtained during training phase capsule pressure - body pressure "conversion table", so as to pilots play an effective protective effect.

For the purpose of this study, we measured the surface pressure of pilot wearing a pressurized service pilot by using a piezoelectric sensor. We selected a total of 4

candidates for the entire experiment. Due to the operation and the software itself Some problems led to the deletion of the experimental data of the first candidate. However, when we still got the better data among the remaining three candidates, Fig. 4 shows that after we use Excel records, we process and analyze the data The actual pressure and various parts of the linear relationship between the measured values.

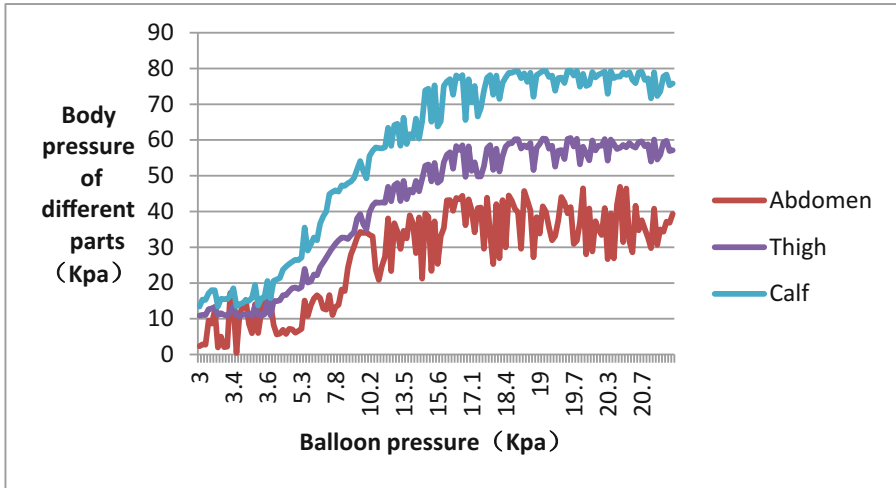


Fig. 4. Subjects’ different parts of the body surface pressure changes with the balloon pressure

As can be seen from Fig. 4, the body surface pressure at 5 sites increases with the increase of capsule pressure. When the capsule pressure increases to a certain extent, the body surface pressure tends to be stable and no longer increases. The reason is that muscle has been compressed to the maximum extent, can not be further compressed, in line with the actual changes in human physiology. Comparisons can be found from various parts can be found, the smallest body surface pressure on the abdomen, calf surface pressure of the largest. This is because the abdominal and thigh parts of the larger force and fat content than smaller legs, resulting in a small surface pressure, in line with the actual situation of human physiology.

Figure 5 depicts the body surface pressure at 3 sites of abdomen, thigh, and lower leg at 5 kPa and 10 kPa balloon pressures. In addition, the error bars for each condition are plotted to show the difference between subjects. It can be seen that the difference of body surface pressure in the abdomen of the subjects is higher than that of the thighs and calves. The reason may be that the abdomen has more soft tissues and the differences among subjects are quite different. Fat content is relatively similar.

Figures 6, 7, 8, 9,10, 11, 12,13, 14, 15, 16, 17, 18, 19 and 20 are based on five parts of the body surface pressure and pressure changes in the relationship between the capsule pressure of three subjects, where the abscissa is the balloon pressure x (in Kpa), the vertical axis is the body surface pressure y (in Kpa), fitted straight line can be used to predict different body surface pressure according to balloon pressure in different parts of the subject. In addition to the poor fitting accuracy of Fig. 6, the other plots’ R2 value are all

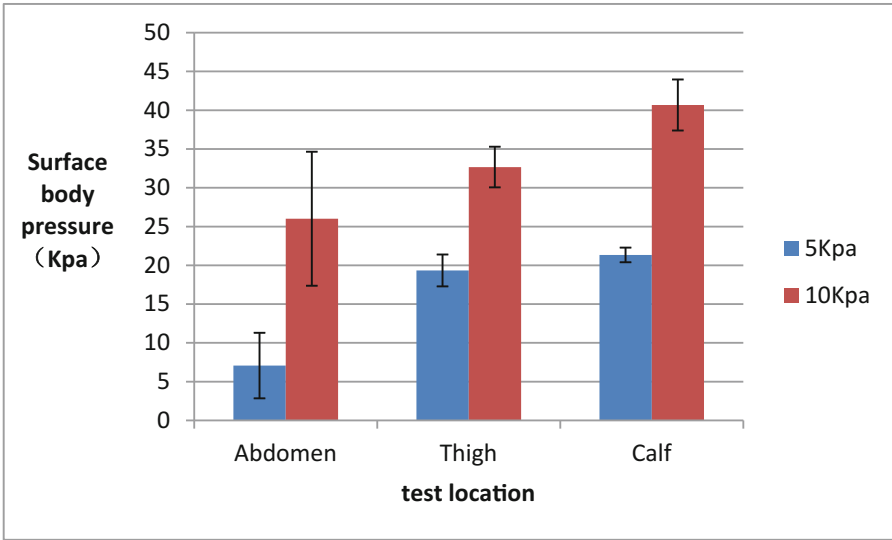


Fig. 5. Different parts and different pressure on the surface of cystic pressure contrast

greater than 0.95, indicating that the fitting accuracy is good and that the surface pressure of different parts of the pilot can be calculated directly from the fitted straight line equation, which will be effective to improve the researcher on the level of anti-Dutch clothing.

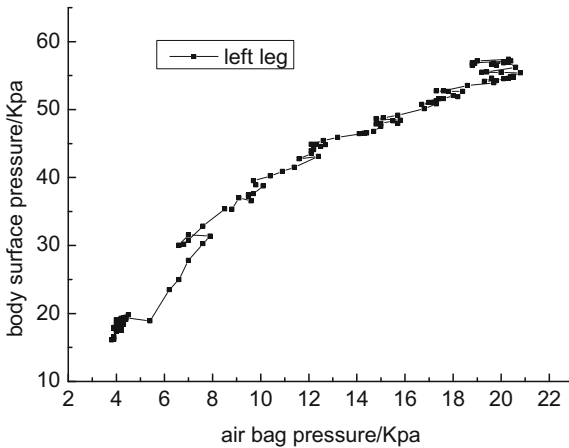


Fig. 6. Left leg body pressure changes with the cyst pressure of subjects 2

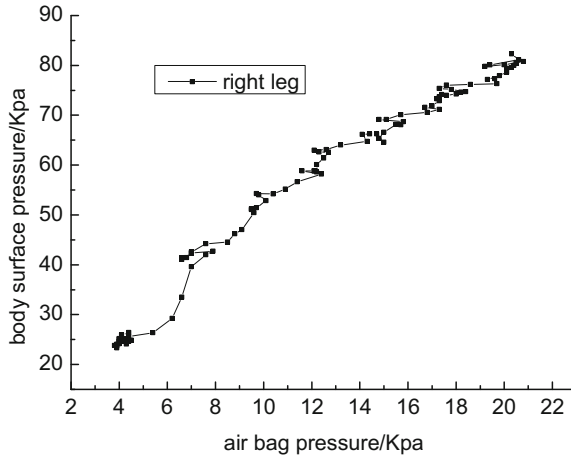


Fig. 7. Right leg body surface pressure changes with the cyst pressure of subjects 2

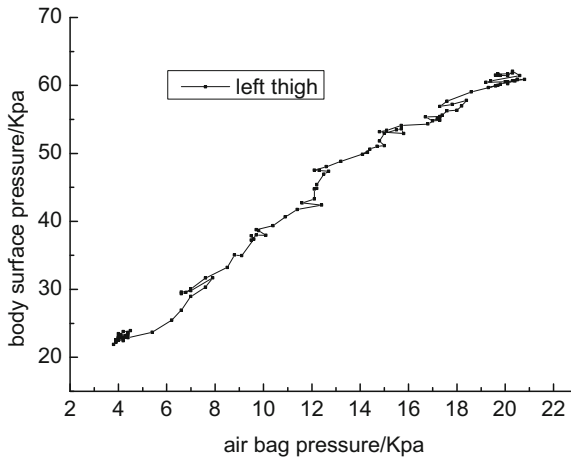


Fig. 8. Left thigh body surface pressure changes with the cyst pressure of subjects 2

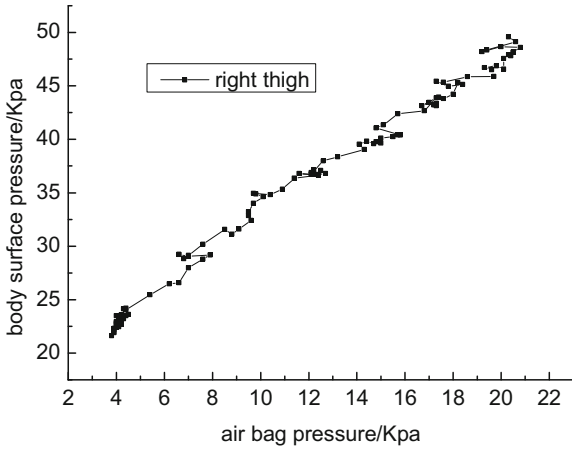


Fig. 9. Right thigh body surface pressure changes with the cyst pressure of subjects 2

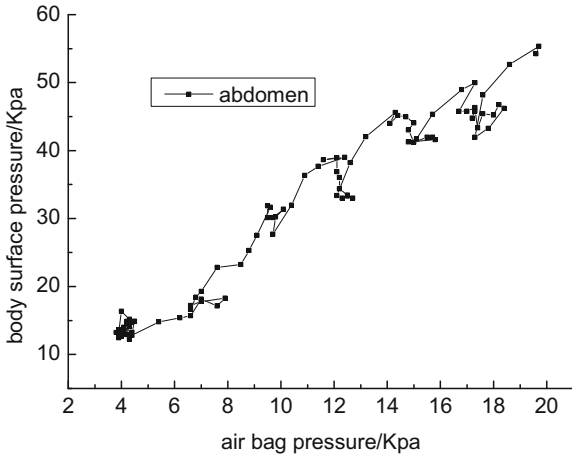


Fig. 10. Abdomen body surface pressure changes with the cyst pressure of subjects 2

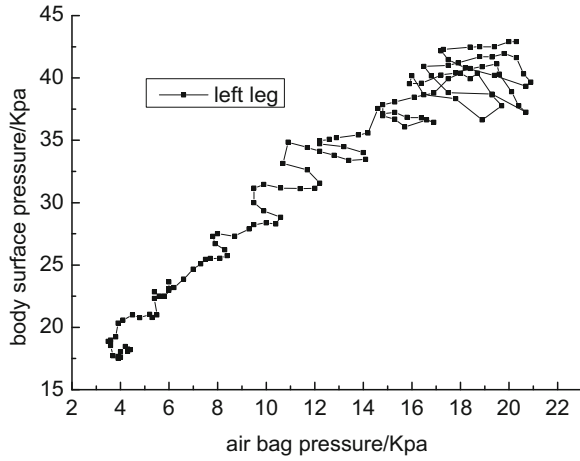


Fig. 11. Left leg surface pressure changes with the cyst pressure of subjects 3

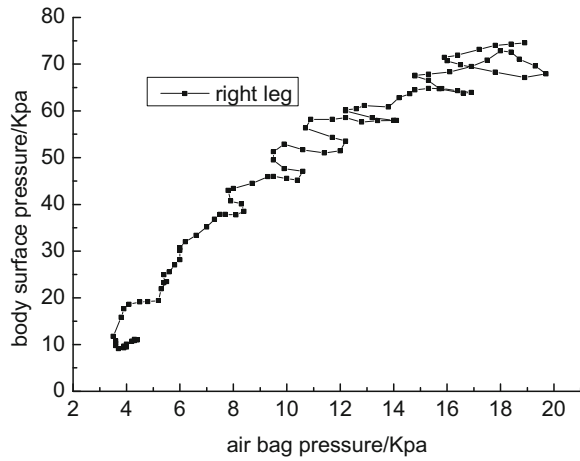


Fig. 12. Right leg surface pressure changes with the cyst pressure of subjects 3

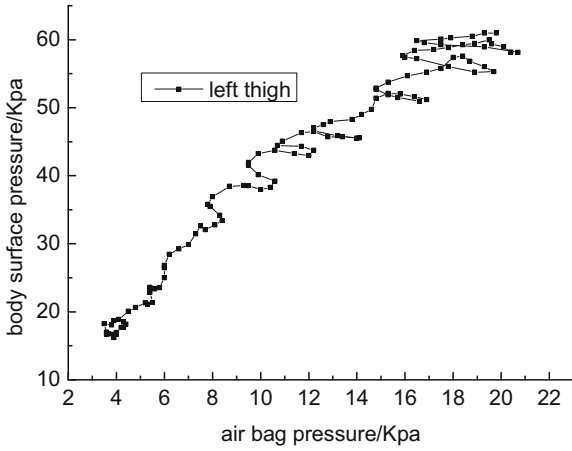


Fig. 13. Left thigh surface pressure changes with the cyst pressure of subjects 3

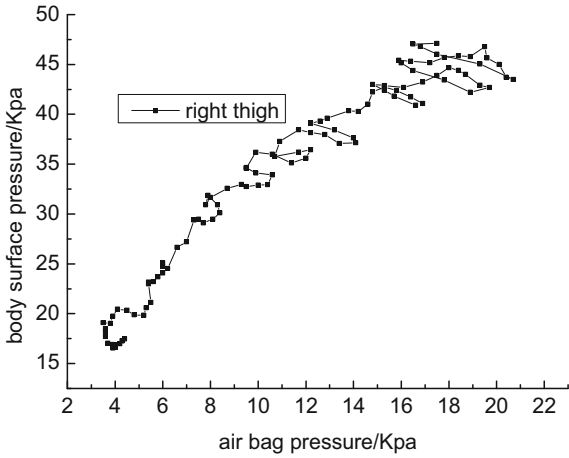


Fig. 14. Right thigh surface pressure changes with the cyst pressure of subjects 3

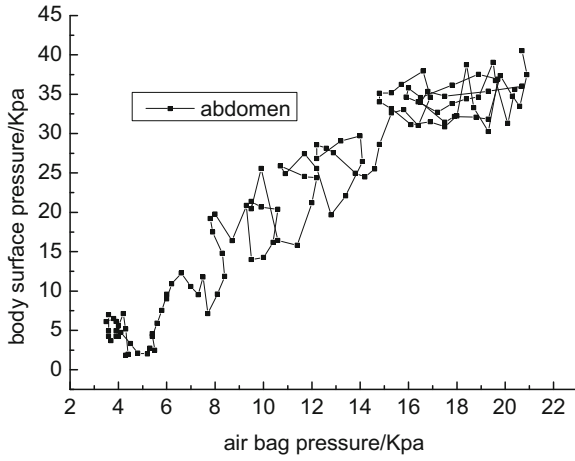


Fig. 15. Abdomen surface pressure changes with the cyst pressure of subjects 3

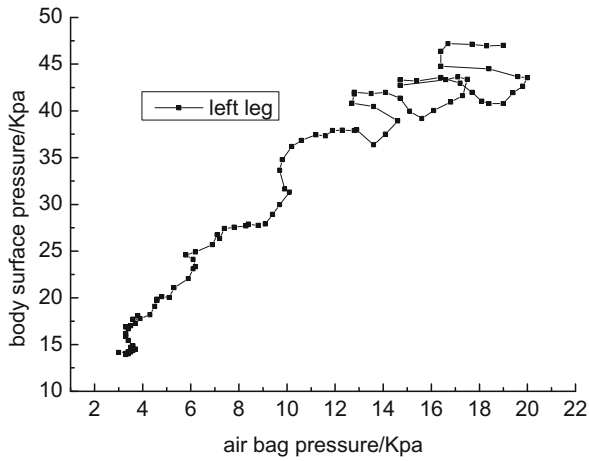


Fig. 16. Left leg surface pressure changes with the cyst pressure of subjects 4

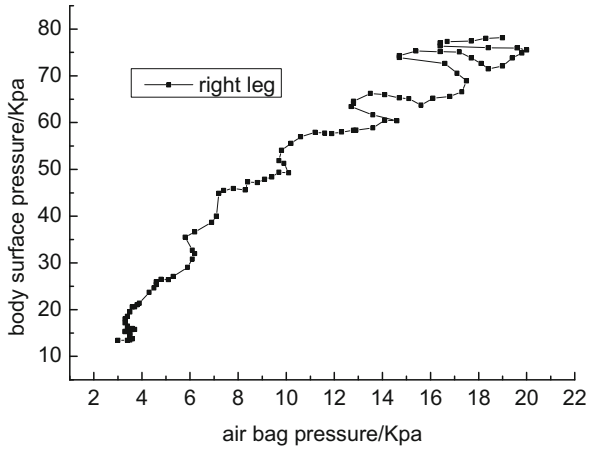


Fig. 17. Right leg body surface pressure changes with the cyst pressure of subjects 4

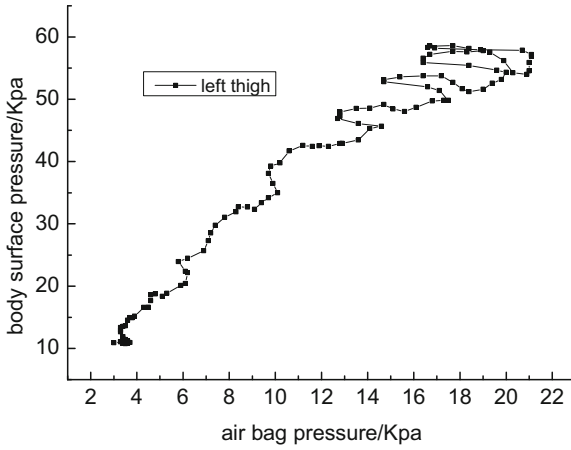


Fig. 18. Left thigh body surface pressure changes with the cyst pressure of subjects 4

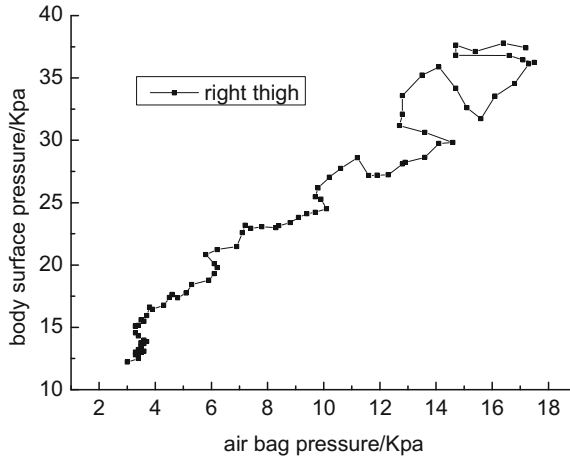


Fig. 19. Right thigh body surface pressure changes with the cyst pressure of subjects 4

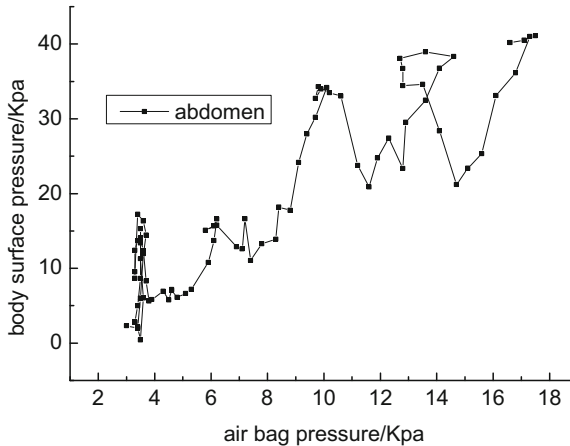


Fig. 20. Abdomen body surface pressure changes with the cyst pressure of subjects 4

3.3 Discussion of the Existing and New Problems

- a. Due to the discrepancy between the experimental equipment and the expectation, we were unable to confirm whether the pressure catcher and the pneumatic pump could be unified at the same time. So we discussed with professor and finally learned that the increase of the pressure pump is set by air pressure of mask. The ultimate mask pressure is controlled by the pilot's breathing rate, and it will be based on a certain proportion of the relationship between the decision to pay the overall service pressure value. Therefore, we finally use the mobile phone to continuously record the data on the screen of the computer to compensate for the pressure changes, to facilitate our follow-up work on the data.

- b. During the experiment, all the data of the first test failed to be fully recorded because of the storage function of the Lab view software itself. When we found out the problem, we found a way to deal with it and remedy it with Professor. And we also came up with the corresponding measures according to the existing problems in advance if similar software problems could be dealt with in the future.

Acknowledgment. Thanks to our tutor, Teacher Wang, she guided our project “Pilot Decompression Service Balloon Pressure Test Research” and this project obtained the subsidy of Beijing Reserve Talent Key Project. Wang also helped us to contact Beihang University and obtained the honorary title of Zhu from Beijing University of Aeronautics and Astronautics. We also get strong support from researchers and pilots at the Air Force Institute of Medicine.

References

1. Poppen, J.R., Drinker, C.K.: Physiological effects and possible methods of reducing the symptoms produced by rapid changes in the speed and direction of airplanes as measured in actual flight. *J. Appl. Physiol.* **3**(4), 204–215 (1950)
2. Brook, W.H.: The development of the Australian anti-G suit. *Aviat. Space Environ. Med.* **61**(2), 176–182 (1990)
3. Sieker, H.O., Martin, E.E.: A comparative study of two experimental pneumatic anti-G suits and the standard USAF G-4A anti-G suit (1953)
4. Burton, R.R., Prakhurs, M.J., Leverett, S.D.: +Gz protection afforded by standard and preacceleration inflations of the bladder and capstan type Gsuits. *Aerosp. Med.* **44**(5), 488–494 (1973)
5. Hollis, L., Barnhill, E., Perrins, M., et al.: Finite element analysis to investigate variability of MR elastography in the human thigh. *Magn. Reson. Imaging* **43**, 27 (2017)
6. Iwamoto, M., Tamura, A., Furusu, K., et al.: Development of a finite element model of the human lower extremity for analyses of automotive crash injuries. *SAE Technical Papers* (2000)
7. Howell, L.L., Jennings, T.J., Loukoumidis, D.G., et al.: Anti-G suit: US, US 4674479 A (1987)
8. Lu, H., Bai, J., Zhang, L., et al.: A simulation study on +Gz protection afforded by extended coverage anti-G suits. *Space Med. Med. Eng.* **11**(4), 240 (1998)
9. Noble, W.E.: Anti-G suit simulator: US, US 5498161 A (1996)

Interaction, Cognition and Emotion



The Effects of Risk and Role on Users' Anticipated Emotions in Safety-Critical Systems

Yusuf Albayram¹(✉), Mohammad Maifi Hasan Khan¹, Theodore Jensen¹,
Ross Buck², and Emil Coman³

¹ Department of Computer Science and Engineering,
University of Connecticut, Storrs, USA

{yusuf.albayram,maifi.khan,theodore.jensen}@uconn.edu

² Department of Communication,
University of Connecticut, Storrs, USA
ross.buck@uconn.edu

³ Health Disparities Institute, University of Connecticut Health Center,
Hartford, USA
coman@uchc.edu

Abstract. Users of safety-critical systems often need to make risky decisions in real-time. However, current system designs do not sufficiently take users' emotions into account. This lack of consideration may negatively influence a user's decision-making and undermine the effectiveness of such a "human-computer collaboration." In a two-way, 2 (role: operator/system administrator) \times 3 (risk level: high/medium/low) factorial study, we investigated the intensity of 44 emotions anticipated by 296 Mechanical Turk users who imagined being the (1) operator or (2) administrator of a drone system identifying (a) enemies on a battlefield, (b) illegal immigrants or (c) whale pods. Results indicated that risk level had a significant main effect on ratings of negative individualistic and negative prosocial emotions. Participants assigned to the high risk scenario anticipated more intense negative individualistic (e.g., nervous) and negative prosocial (e.g., resentful, lonely) emotions and less intense positive (e.g., happy, proud) emotions than participants assigned to the medium and low risk scenarios. We discuss the implications of our findings for the design of safety-critical systems.

Keywords: Emotions · Human-computer interaction
Decision-making · Risk

1 Introduction

Drone systems are increasingly being used for various purposes such as border patrol, battlefield monitoring, target tracking, and recreational activities. These systems can malfunction due to environmental factors, communication errors, or

hardware and software failures, all of which may cause users to experience strong negative emotions (e.g., anger, anxiety, frustration, regret). Although there is a growing body of research showing that emotions strongly influence decision-making under risk and uncertainty [1–3], current safety-critical system designs do not consider users’ emotions. This is likely to undermine effective decision-making, as strong emotions (e.g., regret, suspicion) can alter users’ cognitive process [4].

While the role of emotion was long thought to be disruptive and contrary to models of decision-making, it is now understood that considering only the rational and cognitive is incomplete [5,6]. For instance, prior work in communication theory and psychology suggests that risky situations involve complex strong emotions (e.g., fear, suspicion, excitement) and that, if forewarned about what emotions to expect (i.e., emotional education), people are less surprised by their emotions [1,4,7]. This can allow for mindful processing of risks (e.g., emotional inoculation) [6]. Because of the risks faced by safety-critical system users, we argue that “emotional inoculation” is widely applicable to safety-critical human-computer interaction, and should be explicitly considered while designing user interfaces. A system that communicates about emotions can improve decision-making by allowing users to process the strong negative and positive emotions that arise in their safety-critical tasks. Before designing such systems, it is important to identify the relevant emotions.

As a first step towards this goal, we investigated the effect of risk level and role on users’ anticipated emotions in a two-way, 2 (role: operator/system administrator) \times 3 (risk level: high/medium/low) factorial experiment. We recruited 296 participants on Amazon’s Mechanical Turk platform and provided them with a written description of one of six hypothetical scenarios where they were asked to imagine themselves as a drone operator or system administrator in a high, medium, or low risk scenario. Participants rated the anticipated intensity of 44 emotions in their scenario. Our findings show that risk level had a significant main effect on negative individualistic emotions and negative prosocial emotions. Participants in the high risk scenario expected more negative individualistic (e.g., nervous), more negative prosocial (e.g., resentful, lonely) and fewer positive (e.g., happy) emotions than participants in the medium and low risk scenarios. Insights gained in this study can enhance our understanding of the emotional aspects of decision-making in safety-critical human-computer interaction. The details of our study are presented in the following sections.

2 Background

2.1 Emotions in Decision-Making

Decision-making is the process of selecting a preferred option or course of action among a number of choices [8]. For a considerable time, decision-making was regarded by researchers as a predominantly cognitive process. According to utility theory, decision-makers evaluate the potential consequences of their options and choose the one they believe will yield the most beneficial result (i.e., the

“utility-maximizing” alternative) [9]. Research on decision-making in the last couple of decades has shown that this view is incomplete. There is now a significant amount of psychological research demonstrating that emotions influence decision-making in various ways [4,7].

In a review of these works, Loewenstein and Lerner [4] note two different ways in which emotions enter into a decision: (1) *expected emotions* and (2) *immediate emotions*. *Expected emotions* are those that a decision-maker thinks they will experience as a consequence of some decision. Considered alongside the utility model, the decision-maker will evaluate the consequences of their options and choose that which they expect to maximize positive emotions and minimize negative emotions. *Immediate emotions* are those experienced at the time of decision-making.

Prior work suggests that *immediate emotions* and *expected emotions* are interconnected: *immediate emotions* can impact expectations about future emotions, while *expected emotions* that are anticipated by a decision-maker can influence their current emotional state [4]. For instance, studies have shown that if a decision-maker is presently experiencing positive emotions, his or her evaluation of certain options is likely to be more positive, while those experiencing negative emotions are likely to make more negative evaluations [10,11]. This is exemplified by a “hot/cold empathy gap,” in which individuals in a “hot” emotional state (e.g., angry) have been observed to poorly predict their feelings or behavior when in a “cold” state (e.g., not angry) [12]. Additionally, findings that positive emotions broaden attentional focus while negative emotions narrow it [13,14] suggest that the valence and nature of an individual’s *immediate emotions* influence their cognitive processing. These dynamics have clear implications for decision-making.

In situations involving risk and uncertainty, not only is there a potential increase in cognitive workload, but the effects of the decision-maker’s emotions become more pronounced [1,3]. The “*risk as feelings hypothesis*” explores this notion to explain behavioral responses that differ from what individuals cognitively view as the best course of action. While moderately intense emotions tend to play an “advisory role,” and their influence on an individual’s judgment can often be limited [4,15], strong emotions generally exert more control over behavior. The “*risk as feelings hypothesis*” lends this to the role of “anticipatory” emotions such as fear, worry, and anxiety as inputs in the decision-making process. Specifically, there are a different set of determinants for cognitive evaluations of risk and emotional reactions to risks. While the former is influenced by factors such as outcome probability and severity, emotions are influenced more so by the vividness of imagined consequences or experience with certain outcomes. For instance, feelings about risk have been found to be insensitive to changes in probability, contrary to cognitive evaluations of risk [1].

Use of safety-critical systems is a high-risk, decision-making context where both moderate, advisory emotions and stronger emotions are likely to be at play.

2.2 Emotions in Human-Computer Interaction

Safety-critical system users such as drone operators and air traffic controllers often need to make decisions under uncertainty and time pressure. As wrong decisions may lead to serious consequences for people, property and the environment [16], users of such systems are likely to experience strong anticipatory emotions. Likewise, although the probability of the computer system failing is likely to be low, the potential negative consequences can be emotionally salient. Therefore, it is important to understand what specific emotions may be experienced by users.

Interaction with computers is often portrayed as a purely cognitive endeavor, given that the machines literally operate based on logic. However, recent research highlights the importance of emotional considerations in human computer interaction, wherein a computer that can recognize human emotion can appropriately respond its user's emotions, thus improving the user experience and outcomes of the interaction [17–20]. In one application, Jones and Jonsson [21] proposed an emotionally responsive car system that tracks the emotional state of a driver based on their speech. This information is then used to modify the car's navigational voice, which can relax a tense driver or make them happier about the current conditions. This can improve the driver's concentration and improve safety. This study reports promising results on the potential for emotions to be actively and effectively leveraged in safety-critical human-computer interaction.

Recently, Buck et al. [22] presented the User Affective eXperience (UAX) scale, measuring self-reported emotions that were anticipated in response to pop-up software update messages. They reported 4 latent factors (positive affect, anxiety, hostility and loneliness) which were found to be significantly different between a pressured condition (imagining working on an urgent and stressful task) and a relaxed condition (imagining surfing on the Web while relaxing). Their findings suggest that considering only emotional valence is inadequate, while distinguishing between individualist and pro-social emotions can paint a more thorough picture of the dynamics of affect-influenced decision-making in HCI.

It is fairly obvious that the stress associated with risky, safety-critical system use may cause a user to experience individualistic emotions such as anger or confusion. It is less clear for prosocial emotions, such as guilt and shame, which are those associated with adherence to social norms and group cooperation [23]. First, these are relevant in the drone context because of the presence of other people: system use can have direct consequences for people on the ground, while human operators and administrators work together on tasks with the system. Yet further, a substantial amount of research showing that humans respond socially to computer interaction partners [24,25] suggests that prosocial emotions may arise in the “group cooperation” between human members of the team and the computer system itself. Whereas Freedy et al. [26] sought to define better performance metrics for the unique “interaction of two cognitive systems” (i.e., the human and the computer), we argue that human emotions play an equally important role in the dynamics of such a “collaborative mixed initiative system”.

For example, user emotions may contribute to their “trust” in an automated system, which has been found to influence reliance decisions [27]. Problems of automation *disuse*, in which operators do not use a system when it may help, and *misuse*, in which operators use a system when it is insufficient for some task, are well cited and have been linked to poor “calibration” of trust by the user (“undertrust” and “overtrust,” respectively) [28]. Thus, several researchers have investigated the factors that influence a trust in automation, often varying system reliability and measuring trust with self-reports [29]. While it has been noted that there may be affective components of trust in addition to analogical ones, the role of emotions in trust decisions has not been sufficiently studied. Given that the consequences to poor trust calibration may be particularly severe with safety-critical systems, we argue that affective trust is highly influential on users' decision-making.

While some research efforts have investigated the influence of emotions in human-computer interaction (HCI), to the best of our knowledge, we are the first to investigate the effects of risk and role on users' anticipated emotions in the context of safety-critical drone applications. Specifically, this study expands upon Buck et al.'s work [22] and explores the anticipated intensity of 44 discrete emotions across various roles and risk levels with respect to a safety-critical drone system.

3 Methodology

3.1 Study Design

This study investigates how a safety-critical system users' anticipated emotions vary depending on their role and the criticality of the situation. Toward that, we designed six hypothetical scenarios involving drone operations. Among multiple possible safety critical technologies (e.g., smart grid, self-driving car, assisted robots, drones), this study uses drone because they are utilized for diverse applications (e.g., purely entertainment, border patrol, war).

The experiment was a 2 (role: operator/system administrator) \times 3 (risk level: high/medium/low), between-subject factorial design where participants were randomly assigned to one of six hypothetical scenarios. Participants were asked to rate the anticipated intensity of 44 emotions while imagining themselves in their “risk level” and “role.”

The two “roles” used in the study are as follows:

- **System Administrator:** The task involves managing a drone that is used by someone else (e.g., operator), and making sure the system is working/operating properly.
- **System Operator:** The task involves making decisions with and operating a drone that is overseen by system administrators.

The three “risk levels” used in the study are as follows:

- **High Risk:** The drone was over a battlefield, and the decisions involve identifying enemy targets who may be innocent civilians.
- **Medium Risk:** The drone was over a border region, and the decisions involve arresting suspected illegal immigrants who may be innocent citizens.
- **Low Risk:** The drone was over the ocean, and the decisions involve identifying whale pods or non-interesting seals for a company.

The written descriptions of the scenarios were identical with the exception of the roles and risk level they mentioned, and are outlined in the Appendix. In particular, the hypothetical drone system had some operational instabilities that could cause negative performance. This information was intended to stimulate participants’ emotional responses as they imagined making decisions in a safety-critical situation (i.e., with potentially dangerous consequences) with this imperfect system.

3.2 Survey

We designed a survey consisting of multiple parts as follows.

First, participants were asked to answer demographic questions (e.g., age, gender, and level of education) and report their level of computer proficiency. They were then shown a video about drones and their various applications. Following the video, participants were asked if they understood what drones are, and whether they had prior experience with drones (for either fun or professional reasons).

Subsequently, participants were randomly assigned to one of the six scenarios and, as an attention check, were asked to provide a written explanation of how the drone system is operated, how reliable it is, what their role and task was in the given scenario, and the risks associated with decisions they would have to make.

Finally, participants were asked to rate the expected intensity of 44 different emotions on a scale ranging from 1 (the least amount of intensity) to 7 (the greatest amount of intensity). The emotions were presented in the format “*I would feel [Emotion]*” and shown to participants in random order to avoid biasing them. These emotions were chosen to cover the broad range of emotional responses one could have while using a computer system [22,30,31]. The list of the 44 emotions can be seen in Table 3 in the Appendix.

We expected participants in the high risk scenario (i.e., identifying enemies on a battlefield) to report higher levels of negative emotions (e.g., nervous, anxious) than those in the medium risk (i.e., identifying illegal immigrants) and low risk (i.e., identifying whale pods) scenarios. Additionally, we expected the intensity of negative and positive emotions to vary between operator and administrator roles in the same scenario due to different responsibilities.

Moreover, prior work has found distinction between individualistic and prosocial emotions in response to pop-up software update warning messages [22]. In

our hypothetical context, the distinction between individualistic and prosocial emotions may also be salient, given that (1) system failure could lead to negative consequences for other people and (2) the task involves collaboration with other people and the computer system itself. Thus, we expected to find differences in individualistic and prosocial emotions across risk levels and roles.

3.3 Participants

We recruited participants from Amazon's Mechanical Turk (MTurk) platform. We restricted participants to those 18 or older, currently living in the United States, having greater than 1000 approved HIT's (Human Intelligence Tasks), and having a HIT approval rate greater than 95%.

A total of 300 participants were recruited. We removed the responses of 4 participants who failed to properly answer the attention check question. Thus, a total of 296 valid responses were included in our analysis. Table 1 shows the distributions of participants among the six groups.

Table 1. 6 hypothetical scenarios: 2 roles (i.e., administrator and operator) and 3 risk levels (i.e., high, medium and low risk). The number of participants in each group is also shown.

| | Number of participants | Role | Risk level |
|------------|------------------------|-----------------|-------------|
| Scenario-1 | 48 | System admin | High risk |
| Scenario-2 | 51 | System operator | |
| Scenario-3 | 49 | System admin | Medium risk |
| Scenario-4 | 49 | System operator | |
| Scenario-5 | 50 | System admin | Low risk |
| Scenario-6 | 49 | System operator | |

Participants took an average of 17.7 min (*Median*=14.8, *SD*=11.6 min) to complete the survey and were compensated with \$3. The study was approved by the University's Institutional Review Board (IRB).

3.4 Demographics

Out of 296 participants who completed the survey, 158 (53.4%) were male. Participants' age ranged from 19 to 67 with an average of 33.5 years (median = 32, std = 9.4). All but 3 participants reported English as their native language.

In terms of education level, 89.8% of participants reported having some form of postsecondary education (e.g., college or university) while the most frequent reported education level was a 4-year college degree 43.2% (128). The breakdown of the other reported education levels is as follows: high school/GED (10.1%; 30), some college (23%; 68), 2 year college (14.9%; 44), master's degree (6.4%; 19), and doctoral or professional degree (2.4%; 7).

In terms of reported knowledge about computers in general, 9 (3.0%) participants identified themselves as “beginner,” 5 (1.7%) as “novice,” 90 (30.4%) as “competent,” 150 (50.7%) as “proficient,” and 42 (14.2%) as “expert.” Moreover, 7 (2.4%) participants reported that they did not know what drones were before watching the video, while only one participant reported not knowing after watching the video. Overall, 39 (13.2%) participants reported having had experience with drones for either fun or professional reasons.

To examine demographic differences among the six groups, we performed an exploratory analysis with gender, age, level of education, knowledge about computers, and prior experience with drones. The results of the analysis revealed no significant differences in gender ($\chi^2(5) = 5.79, p = 0.32$), age ($\chi^2(5) = 4.93, p = 0.42$), education ($\chi^2(5) = 6.28, p = 0.27$), reported computer expertise ($\chi^2(5) = 7.86, p = 0.16$) or prior experience with drones ($\chi^2(5) = 5.12, p = 0.40$) across the six groups.

Based on our analysis, we concluded that the groups recruited were similar in terms of demographics.

4 Findings

We first performed an exploratory Principal Component Analysis (PCA) on the ratings of the 44 anticipated emotions. This analysis allowed us to cluster the emotions into groups (i.e., factors) and determine the characteristics of each. Subsequently, for each factor extracted, we performed a 2-way, 2×3 (role \times risk level) Analysis of Variance (ANOVA). The details are presented below.

4.1 Factor Analysis

To assess the appropriateness of the collected emotion data for factor analysis, we first conducted several diagnostic tests using well-known sampling adequacy measures. Bartlett’s test of sphericity measure is ($\chi^2(946) = 8725.2, p < 0.0001$) and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is 0.934. According to the Kaiser criterion, 0.9 and above reveals marvelous value [32], suggesting that our data was correlated and that the variability can be explained by common factors.

Subsequently, we conducted an exploratory PCA on the ratings of the 44 emotions and extracted 6 emotion factors based on the Kaiser criterion (i.e., K1 rule: retain factors if eigenvalue is greater than 1). However, as the Kaiser criterion often leads to substantial overfactoring [33], we also performed parallel analysis and determined the optimal coordinates. Briefly, parallel analysis calculates eigenvalues based on the same sample size and number of variables using sets of random data. Then, each *ith* eigenvalue obtained from the random data is compared with the *ith* eigenvalue produced by the actual data. Based on this comparison, the eigenvalue is retained if the eigenvalue expected from random data is greater than the eigenvalue calculated by the factor analysis. The optimal coordinate method uses linear regression to determine the coordinates where an

eigenvalue diverges [34]. These two methods (i.e., parallel analysis and optimal coordinates) are widely used for determining the appropriate number of factors.

As shown in Fig. 1, both parallel analysis and optimal coordinates suggest extracting three factors for our data. Based on the aforementioned methods, we extracted three factors. These three factors predicted a cumulative total of 55.78% of the variance where factors 1, 2 and 3 explain 29.51%, 18.97%, and 7.29% of the variance, respectively.

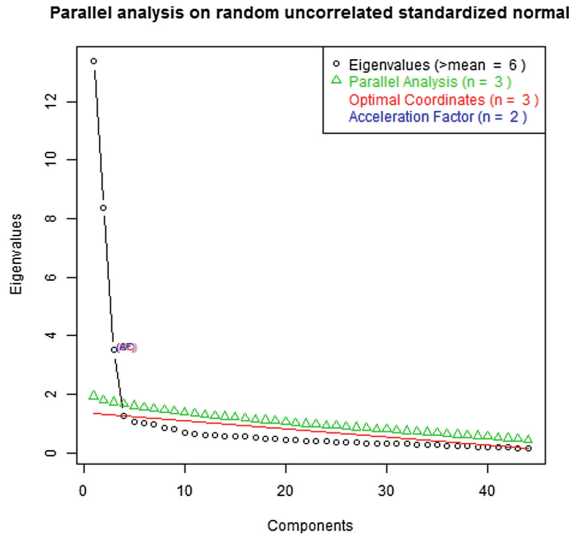


Fig. 1. Scree plot showing eigenvalues from the factor analysis, parallel analysis, optimal coordinates, and acceleration factor.

We used Varimax (orthogonal) as the rotation method, wherein prior work has considered items with a loading above 0.4 to be loaded on a factor [35]. Table 2 shows the rotated factor loadings of 44 emotions as well as the emotions belonging to each factor. Nineteen emotions such as angry, nervous and dismayed were included in Factor-1, which was labeled as “Negative individualistic” emotions. Fifteen emotions such as happy, welcomed and grateful were included in Factor-2, which was labeled as “Positive” emotions. Lastly, ten emotions such as scornful, disdainful and resentful were included in Factor-3, which was labeled as “Negative prosocial” emotions. These factors support those found in Buck et al.’s work [22] in the context of software update pop-up warnings, with our “Negative individualistic” corresponding to their “Anxious,” our “Positive” to that of the same label, and “Negative prosocial” to the pair of factors “Lonely” and “Hostile.”

For our three extracted factors, we also computed reliability measures using Cronbach’s α . As shown in the second to last row of Table 2, all Cronbach’s α values are higher than 0.7. According to McKinley et al. [36], $\alpha > 0.6$ indicates

Table 2. Factor loadings of the 44 emotions from the factor analysis. The highest factor loadings of each factor are highlighted in bold to facilitate visualization. The reliability measures (Cronbach's α and average inter-item correlation (IIC)) are also shown in the last two rows.

| | Factor-1 | Factor-2 | Factor-3 |
|-------------------------------|-------------|-------------|-------------|
| Angry | .775 | | |
| Nervous | .773 | | |
| Dismayed | .771 | | |
| Anxious | .768 | | |
| Distraught | .751 | | |
| Ashamed | .746 | | |
| Down | .732 | | |
| Embarrassed | .731 | | |
| Afraid | .731 | | |
| Guilty | .716 | | |
| Sad | .697 | | |
| Freaked out | .664 | | |
| Depressed | .649 | | |
| Disgusted | .632 | | |
| Confused | .592 | | |
| Dazed | .587 | | |
| Hostile | .528 | .475 | |
| Isolated | .514 | .484 | |
| Surprised | .444 | | |
| Happy | | .776 | |
| Welcomed | | .769 | |
| Grateful | | .762 | |
| Admiring | | .760 | |
| Proud | | .758 | |
| Triumphant | | .758 | |
| Powerful | | .756 | |
| Secure | | .746 | |
| Trusting | | .744 | |
| Friendly | | .737 | |
| Cared-for | | .730 | |
| Respectful | | .717 | |
| Confident | | .681 | |
| Vigorous | | .672 | |
| Energetic | | .668 | |
| Scornful | | | .789 |
| Disdainful | | | .745 |
| Resentful | | | .729 |
| Dishonored | | | .716 |
| Contemptuous | | | .709 |
| Humiliated | | | .669 |
| Arrogant | | | .651 |
| Lonely | | | .602 |
| Insulted | .493 | | .571 |
| Abandoned | .504 | | .515 |
| Cronbach's alpha (α) | .946 | .940 | .886 |
| IIC | .479 | .511 | .525 |

satisfactory internal reliability for all sub-scales. Finally, we calculated average inter-item correlation (IIC) values. As shown in the last row of Table 2, all the sub-scales are above 0.30, indicating “exemplary” reliability [37]. Based on our analysis, we concluded that each of the extracted factors had high reliability.

4.2 ANOVA Analysis

As we wanted to better understand how users might feel while using the drone system in different scenarios and roles, we performed a two-way, (2 × 3) ANOVA for each emotion factor extracted from the factor analysis. More specifically, the dependent variables for our three ANOVAs were negative individualistic emotions (factor-1), positive emotions (factor-2), and negative prosocial emotions (factor-3). We included risk level (high, medium, and low risk), role (system operator and system administrator), and their interaction effects as independent variables in each analysis. The details are presented below.

The ANOVA revealed that risk level had a significant main effect on negative individualistic emotions $F(2,290) = 6.8, p = .001$ and negative prosocial emotions $F(2,290) = 4.1, p = .017$. Participants assigned to the high risk scenario anticipated stronger negative individualistic emotions (e.g., nervous, confused) and negative prosocial emotions (e.g., resentful, lonely), but weaker positive (e.g., happy, grateful) emotions than those assigned to the medium risk and low risk scenarios. More specifically, participants in the high risk scenario ($Mean = 3.54, SD = 1.27$) rated higher negative individualistic emotions than participants in the medium scenario ($Mean = 3.11, SD = 1.32$) and the low risk scenario ($Mean = 2.88, SD = 1.26$). A series of post-hoc pairwise comparisons using Bonferroni correction revealed that the difference in ratings between the

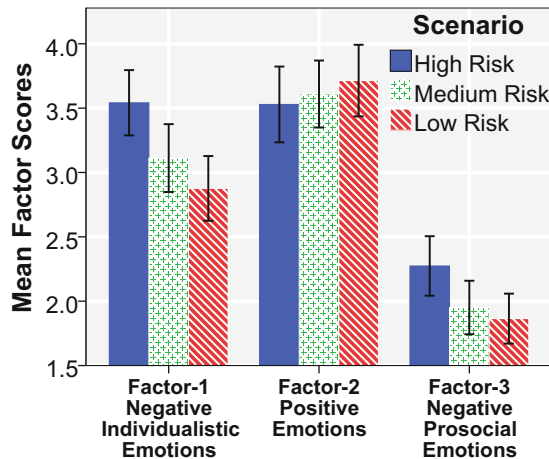


Fig. 2. Mean factor scores for the three risk levels (high risk/medium risk/low risk) for each factor. 95% confidence intervals are also included.

high risk and the low risk scenarios was significant ($p < .001$). Similarly, participants in the high risk scenario ($Mean = 2.27$, $SD = 1.16$) rated negative prosocial emotions higher than participants in the medium risk scenario ($Mean = 1.95$, $SD = 1.04$) and the low risk scenario ($Mean = 1.86$, $SD = 0.97$). A series of post-hoc pairwise comparisons using Bonferroni correction revealed that the difference in ratings between the high risk and low risk scenarios was significant ($p < .021$). Although those in the high risk scenario ($Mean = 3.53$, $SD = 1.48$) rated lower levels of positive emotions than participants in the medium risk scenario ($Mean = 3.61$, $SD = 1.30$) and low risk scenario ($Mean = 3.71$, $SD = 1.40$), the difference in ratings among the three risk levels was not statistically significant. The mean factor scores for the three risk levels are shown in Fig. 2.

The ANOVA also revealed that there was no significant main effect on emotions due to role. The mean factor scores for the two roles (operator/system administrator) can be seen in Fig. 3.

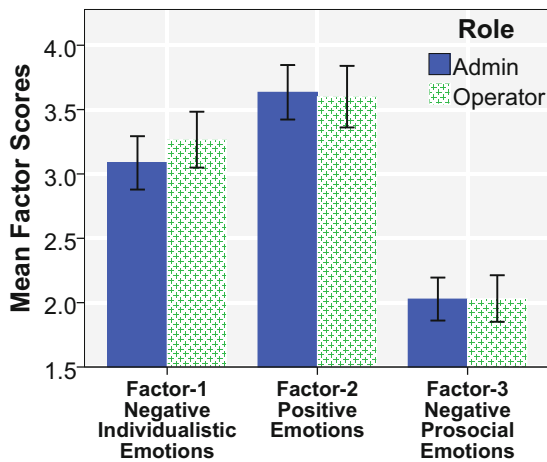


Fig. 3. Mean factor scores for the two roles (system administrator/system operator) for each factor. 95% confidence intervals are also included.

5 Discussion

Despite a growing body of literature demonstrating the significant role of emotions in the decision-making process, we have a relatively limited understanding of the specific emotions relevant to high risk decision-making. As safety-critical technologies such as drones and self-driving cars become more prevalent, so will the high-risk decisions to which their users must attend. To gain insight into the effect of risk level and role on safety-critical system users' emotions, we asked participants to imagine themselves as a drone operator or system administrator

in a high, medium, or low risk scenario. They then rated the expected intensity of 44 emotions while imagining the scenario.

We found that participants in the high risk scenario reported higher levels of negative individualistic emotions (e.g., angry, nervous), negative prosocial emotions (e.g., scornful, resentful) and lower levels of positive emotions (e.g., happy, grateful) than participants assigned to the medium and low risk scenarios. These differences were significant between high and low risk participants for both negative prosocial and negative individualistic emotions. These findings suggest that, unsurprisingly, use of safety-critical systems may involve strong negative emotions. The notion that computers are cognitive entities, with which interaction should be non-emotional in order to be efficient and successful, may be particularly destructive in this context. A lack of acknowledgment by the system may not only alter a user's decision-making, but lead to stronger negative emotions that impact later interaction.

Developing emotion-aware communication strategies by detecting users' emotions during system operations can reduce the potentially harmful effects of negative emotions. Specifically, teaching users to recognize their emotions (emotional education) may enable them to act more mindfully, and help to lessen the potential negative effects of strong emotions on decision-making (emotional inoculation) [6]. We argue that "emotional inoculation" is particularly applicable in the safety-critical domain, such as our hypothetical drone system. Communicating with users about emotions they may experience while using a system can positively contribute to both their decision-making outcomes and their perceptions of the system. Future work should test the effectiveness of safety-critical system interfaces that incorporate emotional inoculation via different types of messages and in various decision-making contexts. Furthermore, "emotional inoculation" and "emotional education" can be incorporated into training materials for safety-critical system users (e.g., drone operators). Using virtual simulators in realistic scenarios, such training systems could inform operators about the emotions they might experience during certain points of system use (e.g., feeling nervous and anxious during a time-sensitive task) and the nature of the specific emotions in such situations (e.g., prosocial vs. individualistic, or positive vs. negative). This can help prepare operators to regulate their reactions under time pressure and stress while performing complex safety-critical tasks [38].

These kinds of emotional communication can help to improve a user's trust calibration. Prior work has found that happiness, as well as "liking" a system influence reliance [39]. These affective aspects may help to explain changes in trust over the course of a human-computer interaction [29,39,40]. Future work should explore how the negative individualistic and negative prosocial emotions associated with safety-critical system use factor into trust evaluations and reliance decisions, as well as how an understanding of these emotions can be leveraged to improve system design and trust calibration.

We also found that, at the same risk level, the intensity of emotion factors differed (see Figs. 2 and 3). Negative prosocial emotions had the lowest mean intensity in all risk levels and roles, whereas positive emotions and negative individualistic emotions generally had higher intensity. Though prosocial emotions were not felt as strongly by participants, we observed that their anticipated intensity differed between high and low risk level participants. It appears that users are not just thinking about themselves with their use of the drone system, but about the involvement of others. This result is in line with research demonstrating the relevance of both individualistic and prosocial emotions in the context of pop-up security messages [22]. In the drone context, prosocial emotions could have been associated with (1) people on the ground who may have been impacted by the drone, (2) other human collaborators, or (3) the computer system itself. The latter is supported by research demonstrating social responses to computers by human users [24]. Future work could shed light on the specific effect that the computer itself has on user emotions by investigating how factors of the system and its interface influence the intensity of prosocial emotions, relative to differences in the context of system use.

Lastly, we found that for all the three factors, the interaction between risk level and role was not significant. This indicates that participants' emotions were more likely to be influenced by the criticality of the situation rather than their assigned role. It is possible that participants in operator and administrator roles in the same scenario considered the level of risk the same, and thus the role to which they were assigned did not make a strong contribution to their overall feelings. Such a difference may be more pronounced in a lab setting where participants interact directly with a system. If the user's role on a task-oriented team is more linked to their actions, then emotions may be impacted by their level of responsibility for team success.

5.1 Limitations

While this study provides insights about the effects of risk and role on users' emotions, there are several limitations in this work.

First, we used hypothetical (i.e., artificial) scenarios in which participants rated how they would expect to feel as the operator or administrator of a drone system. Given the lack of actual interaction with a computer system, it may have been difficult for participants to anticipate the emotions they would experience. Moreover, this could contribute to misinterpretations of the degree of risk. For example, some participants in the low risk condition (i.e., identifying whale pods) may have considered the situation to be very risky, since failure could have caused "job loss." Nevertheless, even in this artificial scenario-based methodology, our results revealed considerably diverse ratings of emotions depending on the group to which participants were assigned.

Second, we recruited participants from the MTurk platform. Although MTurk allows for recruiting larger and more diverse populations in terms of age, education level and ethnicity compared to samples from specific subpopulations (e.g., students enrolled in a psychology class) [41, 42], it is hard to verify the attentiveness of MTurk users. To filter out responses that demonstrated a lack of understanding of the scenario, we included an attention check question in the survey.

Lastly, since our study was survey-based, emotional states of participants were measured via self-reports. Though our data provides insight into the role of “anticipated emotions” in a risky human-computer interaction, it needs further validation given that individuals may have difficulty predicting their emotional states [43]. To develop a more thorough understanding of user’s emotions, future studies should investigate the somatic components (e.g., facial expressions and the heartbeat) [44] of “immediate emotions” in studies involving actual human interaction with a computer system.

We believe that this work is a useful starting point for research on the role of emotions in decision-making with safety-critical systems, which has important implications for system interface design. We encourage future work to investigate the specific factors that influence user emotions (e.g., risk and the nature of consequences, organizational structure, system features) as well as the influence that different types of emotions have on decision-making, behavior, and system performance.

6 Conclusion

This study aimed to understand the role of emotions in decisions at various risk levels and responsibilities with respect to a safety-critical system. Participants were asked to rate the intensity with which they would feel 44 emotions while imagining using a drone system in one of six hypothetical scenarios where they were asked to imagine themselves as a drone operator or system administrator in a high, medium, or low risk scenario. We found that participants assigned to the high risk scenario anticipated more intense negative individualistic, negative prosocial and less intense positive emotions than participants assigned to medium and low risk scenarios. We strongly believe that insights gained in this work will enable researchers to develop more effective emotionally-aware communication strategies for safety-critical systems.

Acknowledgments. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-15-1-0490.

Appendix

The Descriptions of the Scenarios

Through a two-way 2 (role: system operator/system administrator) \times 3 (risk level: high risk/medium risk/low risk) factorial design experiment, participants were assigned to one of the six hypothetical scenarios. Depending on the assigned role and risk level, participants were shown one of the phrasings separated by vertical bars (|) below. For instance, participants assigned to the system operator role in the high risk scenario were shown (*Role_{opr}*) and (*Risk_{high}*), while participants assigned to the system administrator role in the low risk scenario were shown (*Role_{adm}*) and (*Risk_{low}*), and so on. The entire written description of scenarios are outlined below.

Now, imagine that [*Role_{opr}*: *there are system administrators who are*] | [*Role_{adm}*: *you are the system administrator who is*] responsible for:

- Making sure that the software of the system that is used to operate the drone remotely is up-to-date.
- Making sure that the hardware of the system is up-to-date.
- Troubleshooting of the system if the performance is not acceptable.
- Performing preventative maintenance of the system.

However, despite [*Role_{opr}*: *their*] | [*Role_{adm}*: *your*] best effort, the system is not perfectly reliable and the system occasionally experiences the followings due to software bugs/hardware failures:

- The system occasionally crashes due to some unknown reasons and takes 2 min to reboot, making the system unavailable, and the timing/frequency of the crash is unpredictable.
- The system occasionally becomes very slow (e.g., freezes for 10 s at a time) due to unknown software/hardware bugs.
- The system occasionally drops video frames due to communication errors.
- Different hardware components of the system rarely fails (e.g., once every 6 months).

Now, imagine that [*Role_{opr}*: *you are asked to use the system to make*] | [*Role_{adm}*: *the drone that you are responsible for managing is going to be used by someone else (e.g., operator) whose*] decisions involve identifying:

- *Risk_{high}*: enemy targets in a battlefield where there may be also innocent civilians.
- *Risk_{med}*: arresting or not arresting suspected illegal immigrants who may be innocent citizens in a border region.
- *Risk_{low}*: whale pods or non-interesting seals in the ocean for a company.

Table 3. Participants were asked to rate the expected intensity of these 44 emotions on a scale ranging from 1 (the least amount of intensity) to 7 (the greatest amount of intensity).

| Emotions |
|--|
| 1. I would feel TRUSTING (e.g. because the system has given an opportunity to respond) |
| 2. I would feel HAPPY (e.g., because I am informed of actual system states) |
| 3. I would feel CONFIDENT (e.g., because I am informed of actual system states) |
| 4. I would feel SECURE (e.g., because I am informed of actual system states) |
| 5. I would feel SAD (e.g., because the system is not performing as expected) |
| 6. I would feel DEPRESSED (e.g., because the system is not performing as expected) |
| 7. I would feel DOWN (e.g., because the system is not performing as expected) |
| 8. I would feel AFRAID (e.g., because the system is not performing as expected) |
| 9. I would feel NERVOUS (e.g., because the system is not performing as expected) |
| 10. I would feel ANXIOUS (e.g., because the system is not performing as expected) |
| 11. I would feel ANGRY (e.g., because the system is not performing as expected) |
| 12. I would feel INSULTED (e.g., because the system is not performing as expected) |
| 13. I would feel HOSTILE (e.g., because the system is not performing as expected) |
| 14. I would feel SURPRISED (e.g., because one does not expect the interruption) |
| 15. I would feel DAZED (e.g., because one does not expect the interruption) |
| 16. I would feel CONFUSED (e.g., because one does not expect the interruption) |
| 17. I would feel FREAKED OUT (e.g., because one does not expect the interruption) |
| 18. I would feel DISGUSTED (e.g., because the system is not performing as expected) |
| 19. I would feel DISMAYED (e.g., because the system is not performing as expected) |
| 20. I would feel DISTRAUGHT (e.g., because the system is not performing as expected) |
| 21. I would feel CARED-FOR (e.g., because I am informed of actual system states) |
| 22. I would feel FRIENDLY (e.g., because I am informed of actual system states) |
| 23. I would feel WELCOMED (e.g., because I am informed of actual system states) |
| 24. I would feel POWERFUL (e.g., because I am warned and can respond) |
| 25. I would feel ENERGETIC (e.g., because I am warned and can respond) |
| 26. I would feel VIGOROUS (e.g., because I am warned and can respond) |
| 27. I would feel ISOLATED (e.g., because my response may be inadequate) |
| 28. I would feel LONELY (e.g., because my response may be inadequate) |
| 29. I would feel ABANDONED (e.g., because my response may be inadequate) |
| 30. I would feel PROUD (e.g., because I am warned and can respond) |
| 31. I would feel TRIUMPHANT (e.g., because I am warned and can respond) |
| 32. I would feel ARROGANT (e.g., because I am warned and can respond) |
| 33. I would feel ASHAMED (e.g., because my response may be inadequate) |
| 34. I would feel GUILTY (e.g., because my response may be inadequate) |
| 35. I would feel EMBARRASSED (e.g., because my response may be inadequate) |
| 36. I would feel SCORNFUL (e.g., because the system state is fine) |
| 37. I would feel CONTEMPTUOUS (e.g., because the system state is fine) |
| 38. I would feel DISDAINFUL (e.g., because the system state is fine) |
| 39. I would feel HUMILIATED (e.g., because the system state is fine) |
| 40. I would feel DISHONORED (e.g., because the system state is fine) |
| 41. I would feel RESENTFUL (e.g., because the system state is fine) |
| 42. I would feel GRATEFUL (e.g., because the system has given an opportunity to respond) |
| 43. I would feel RESPECTFUL (e.g., because the system has given an opportunity to respond) |
| 44. I would feel ADMIRING (e.g., because the system has given an opportunity to respond) |

References

1. Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N.: Risk as feelings. *Psychol. Bull.* **127**(2), 267 (2001)
2. Schlösser, T., Dunning, D., Fetschenhauer, D.: What a feeling: the role of immediate and anticipated emotions in risky decisions. *J. Behav. Decis. Mak.* **26**(1), 13–30 (2013)
3. Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. *Econom.: J. Econom. Soc.* **47**, 263–291 (1979)
4. Loewenstein, G., Lerner, J.S.: The role of affect in decision making. *Handb. Affect. Sci.* **619**(642), 3 (2003)
5. Buck, R., Davis, W.A.: Marketing risk: emotional appeals can promote the mindless acceptance of risk. In: Roeser, S. (ed.) *Emotions and risky technologies*, pp. 61–80. Springer, Dordrecht (2010). https://doi.org/10.1007/978-90-481-8647-1_4
6. Buck, R., Ferrer, R.: Emotion, warnings, and the ethics of risk communication. In: Roeser, S., Hillerbrand, R., Sandin, P., Peterson, M. (eds.) *Handbook of Risk Theory*, pp. 693–723. Springer, Dordrecht (2012). https://doi.org/10.1007/978-94-007-1433-5_27
7. Lerner, J.S., Li, Y., Valdesolo, P., Kassam, K.S.: Emotion and decision making. *Ann. Rev. Psychol.* **66**, 799–823 (2015)
8. Wilson, R.A., Keil, F.C.: *The MIT Encyclopedia of the Cognitive Sciences*. MIT press, Cambridge (2001)
9. Harless, D.W., Camerer, C.F.: The predictive utility of generalized expected utility theories. *Econom.: J. Econom. Soc.* **62**, 1251–1289 (1994)
10. Clore, G.L.: Cognitive phenomenology: feelings and the construction of judgment. *Const. Soc. Judgm.* **10**, 133–163 (1992)
11. Clore, G.L., Schwarz, N., Conway, M.: Affective causes and consequences of social information processing. *Handb. Soc. Cogn.* **1**, 323–417 (1994)
12. Loewenstein, G.: Out of control: visceral influences on behavior. *Organ. Behav. Hum. Decis. Process.* **65**(3), 272–292 (1996)
13. Basso, M.R., Schefft, B.K., Ris, M.D., Dember, W.N.: Mood and global-local visual processing. *J. Int. Neuropsychol. Soc.* **2**(3), 249–255 (1996)
14. Conway, M., Giannopoulos, C.: Dysphoria and decision making: limited information use for evaluations of multiattribute targets. *J. Pers. Soc. Psychol.* **64**(4), 613 (1993)
15. Forgas, J.P.: Mood and judgment: the affect infusion model (AIM). *Psychol. Bull.* **117**(1), 39 (1995)
16. Knight, J.C.: Safety critical systems: challenges and directions. In: *Proceedings of the 24th International Conference on Software Engineering, ICSE 2002*, pp. 547–550. ACM, New York (2002)
17. Brave, S., Nass, C.: *The Human-Computer Interaction Handbook*, pp. 81–96. L. Erlbaum Associates Inc., Hillsdale (2003)
18. Brave, S., Nass, C., Hutchinson, K.: Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *Int. J. Hum.-Comput. Stud.* **62**(2), 161–178 (2005)
19. Picard, R.W., Klein, J.: Computers that recognise and respond to user emotion: theoretical and practical implications. *Interact. Comput.* **14**(2), 141–169 (2002)
20. Klein, J., Moon, Y., Picard, R.W.: This computer responds to user frustration: theory, design, and results. *Interact. Comput.* **14**(2), 119–140 (2002)

21. Jones, C., Jonsson, I.-M.: Using paralinguistic cues in speech to recognise emotions in older car drivers. In: Peter, C., Beale, R. (eds.) *Affect and Emotion in Human-Computer Interaction*. LNCS, vol. 4868, pp. 229–240. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85099-1_20
22. Buck, R., Khan, M., Fagan, M., Coman, E.: The user affective experience scale: a measure of emotions anticipated in response to pop-up computer warnings. *Int. J. Hum.-Comput. Interact.* **34**, 1–10 (2017)
23. Bowles, S., Gintis, H.: *Prosocial Emotions. The Economy As an Evolving Complex System III*, pp. 339–366. Santa Fe Institute, Santa Fe (2005)
24. Reeves, B., Nass, C.I.: *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, Cambridge (1996)
25. Nass, C., Moon, Y.: Machines and mindlessness: social responses to computers. *J. Soc. Issues* **56**(1), 81–103 (2000)
26. Freedy, A., DeVisser, E., Weltman, G., Coeyman, N.: Measurement of trust in human-robot collaboration. In: *2007 International Symposium on Collaborative Technologies and Systems, CTS 2007*, pp. 106–114. IEEE (2007)
27. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.* **58**(6), 697–718 (2003)
28. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* **39**(2), 230–253 (1997)
29. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. *Hum. Factors* **46**(1), 50–80 (2004)
30. Buck, R., Anderson, E., Chaudhuri, A., Ray, I.: Emotion and reason in persuasion: applying the ari model and the casc scale. *J. Bus. Res.* **57**(6), 647–656 (2004)
31. Kay, R.H., Loverock, S.: Assessing emotions related to learning new software: the computer emotion scale. *Comput. Hum. Behav.* **24**(4), 1605–1623 (2008)
32. Kaiser, H.F.: An index of factorial simplicity. *Psychometrika* **39**(1), 31–36 (1974)
33. Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., Strahan, E.J.: Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* **4**(3), 272 (1999)
34. Raïche, G., Walls, T.A., Magis, D., Riopel, M., Blais, J.G.: Non-graphical solutions for Cattell's scree test. *Methodol.: Eur. J. Res. Methods Behav. Soc. Sci.* **9**(1), 23–29 (2013)
35. Stevens, J.P.: *Applied Multivariate Statistics for the Social Sciences*. Routledge, London (2012)
36. McKinley, R.K., Manku-Scott, T., Hastings, A.M., French, D.P., Baker, R.: Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the united kingdom: development of a patient questionnaire. *BMJ: Br. Med. J.* **314**(7075), 193–198 (1997)
37. Robinson, J.P., Shaver, P.R., Wrightsman, L.S.: Criteria for scale selection and evaluation. *Meas. Personal. Soc. Psychol. Attitudes* **1**(3), 1–16 (1991)
38. Luini, L.P., Marucci, F.S.: Prediction-confirmation hypothesis and affective deflection model to account for split-second decisions and decision-making under pressure of proficient decision-makers. *Cogn. Technol. Work* **17**(3), 329–344 (2015)
39. Merritt, S.M.: Affective processes in human-automation interactions. *Hum. Factors* **53**(4), 356–370 (2011)
40. Merritt, S.M., Ilgen, D.R.: Not all trust is created equal: dispositional and history-based trust in human-automation interactions. *Hum. Factors* **50**(2), 194–210 (2008)

41. Chandler, J.J., Paolacci, G.: Lie for a Dime: when most prescreening responses are honest but most study participants are impostors. *Soc. Psychol. Personal. Sci.* **8**(5), 500–508 (2017)
42. Landers, R.N., Behrend, T.S.: An inconvenient truth: arbitrary distinctions between organizational, mechanical turk, and other convenience samples. *Ind. Organ. Psychol.* **8**(2), 142–164 (2015)
43. Picard, R.W., Picard, R.: *Affective Computing*, vol. 252. MIT press, Cambridge (1997)
44. Han, S., Lerner, J.S., Sander, D., Scherer, K.: Decision making. In: Sander, D., Scherer, K.R. (eds.) *The Oxford Companion to Emotion and the Affective Sciences*. Oxford University Press, Oxford (2009)



Comparison of Intellectus Statistics and Statistical Package for the Social Sciences

Differences in User Performance Based on Presentation of Statistical Data

Allen C. Chen^(✉), Sabrina Moran, Yuting Sun, and Kim-Phuong L. Vu

California State University, Long Beach, Long Beach, CA 90840, USA
achen31@gmail.com, sabrina.n.moran@gmail.com,
debby130403@gmail.com, kim.vu@csulb.edu

Abstract. Data-to-text systems create reports using natural language to simplify the presentation of complex data. Intellectus Statistics (IS) is a cloud-based statistical analysis software that provides users with output displayed in American Psychological Association (APA) narrative format. Statistical Package for the Social Sciences (SPSS) is another statistical analysis software; however, SPSS output is mainly presented numerically with tables and graphs. The purpose of this study was to compare the effectiveness and efficiency of using IS and SPSS to conduct and interpret analyses. An output presented in narrative format could be beneficial to students learning statistics who may have difficulty interpreting results. Overall, accuracy scores and time on task for the two software were not significantly different. Perceived usability and ease of use ratings for IS were significantly higher compared to SPSS. On the other hand, ratings of perceived usefulness were not significantly different between the two software. Results also suggested that participants preferred IS and felt more confident in conducting statistical analyses when using the software. Though there was no significant difference in task accuracy between the two software, data-to-text output helped students with interpreting assumptions for analyses and formatting written results.

Keywords: Data-to-text systems · Data interpretation
Visual display of information

1 Background

Statistics is the science of collecting, analyzing, and interpreting data. To produce accurate reports and interpretations of data, appropriate steps must be taken prior to analyses. Some common pitfalls that lead to inaccurate results include incorrect analysis choice, violations of assumptions, and incorrect interpretation of results. As the demand for statistical knowledge has grown in the age of big data, employment for statisticians is projected to increase 34% between 2014 and 2024 [9]. Furthermore, a report by Wasserstein [10] noted that completion rate for bachelor's degrees in statistics outpaced all other STEM disciplines in the four years prior to the paper's publication.

1.1 Intellectus Statistics

Intellectus Statistics (IS) is a statistical analysis software created by Statistics Solutions. IS provides a cloud-based platform where users are presented with output displayed in APA narrative format after conducting statistical analyses. The data-to-text output also reports on the assumptions of analyses and indicates any violations. Past research has shown that students in introductory statistics courses have difficulty writing conclusions based on results from their data analysis [5]. Data-to-text systems simplify the presentation of data by using natural language to generate reports [3], which may be beneficial to users with limited statistical background.

Potential users of IS include anyone seeking to conduct statistical analyses. With an APA formatted output, the target user group is assumed to be undergraduate and graduate students in the social sciences. The results, which includes graphs and tables, are presented in a format commonly used in those disciplines. The interface was designed to be simple and easy to use, and an output in the form of a narrative would appeal to users with limited statistical background. In addition to conducting different types of analyses, the software also assists with data cleaning and data visualizations.

The ability to access the software and saved data from any computer with an internet connection makes IS attractive to users who rely on public computers or have limited storage space. While being cloud-based has its advantages, the platform is constrained by maintenance of working code, product support, and the version update process. Any potential issues with the cloud-based software would affect all users.

1.2 Statistical Package for the Social Sciences

Statistical Package for the Social Sciences (SPSS) is another statistical analysis software. The original target users of SPSS were students and professors in the social sciences, but its use has grown to other fields such as the medical sciences. SPSS offers greater functionality compared to IS; however, the interface is less intuitive and often uses terms unfamiliar to novice users. Much of the output, presented numerically with tables, requires some statistical background for accurate interpretation.

Potential users of SPSS consist of individuals who need to conduct statistical analyses. SPSS is commonly used by and taught to psychology students at many educational institutions. Through the use of commands in the Syntax Editor, SPSS offers intermediate and expert users increased flexibility and efficiency of use. Another benefit, due to its wide use in academia, is the wealth of help documentation available on the web. SPSS is often daunting for novice users due to the amount of information, options, and unfamiliar terminology. The program can be downloaded and installed on individual computers. Software updates are not automatic and are based on user preference. A cloud-based version of SPSS is also available, but its use could result in the same problems discussed earlier with cloud-based programs.

1.3 Purpose of Current Study

The purpose of this usability evaluation was to compare IS and SPSS' effectiveness and efficiency for students who have some prior experience conducting statistical analyses. An output presented in narrative format could be beneficial to students who have trouble interpreting results by providing more contextual information alongside the numbers. The students in the current study completed tasks involving data entry, data cleaning, conducting statistical analyses, interpretation of assumptions, and interpretation of results. Task accuracy, time on task, and perceived usability was measured and compared. Perceived usability was measured using the System Usability Scale (SUS). Perceived ease of use and usefulness was measured using the extended Technology Acceptance Model (TAM 2).

2 Method

2.1 Metrics

Time on task, task accuracy, and task completion rate were used to assess efficiency and effectiveness. Time on task was calculated by measuring the length of time that participants spent on each subtask. Length of time began when the participant started moving the computer mouse for navigation and ended when the participant verbally confirmed completion of the subtask. Combining the times for each subtask resulted in the completion time for the overall task. Shorter task times would indicate increased efficiency [1].

Task accuracy was determined using a scoring rubric. Points were awarded for correctly running analyses and correct interpretation of output. Scoring for interpretation was also dependent on including necessary information for APA formatted results. Points for originality were awarded based on Turnitin Similarity scores, which indicated a percent of matching content with other sources. The same scoring rubric was used for both software and tasks. Task accuracy was scored by two independent raters and high inter-rater reliability was found (Cronbach's $\alpha = 0.87$ to 0.96). Higher task accuracy would indicate increased task success and increased effectiveness [1].

Overall perceived usability was obtained using the System Usability Scale (SUS). The SUS is a self-reported 10-item Likert scale questionnaire that measures overall perceived reliability. According to previous research, the SUS is a highly robust tool for measuring perceived usability [2]. Possible perceived usability scores range from zero to 100. Scores under 50 indicate unacceptable usability, while scores above 70 indicate acceptable usability [1]. Scores ranging from 51 to 70 indicate marginal acceptability.

Perceived usefulness and ease of use was obtained using scales from the extended Technology Acceptance Model (TAM 2). Perceived usefulness and ease of use were reduced to four Likert-scale items ranging from one to seven in the extended model. Across studies and time periods, the internal reliability for both perceived usefulness (Cronbach's $\alpha = 0.87$ to 0.98) and ease of use (Cronbach's $\alpha = 0.86$ to 0.98) in the TAM 2 were high [8]. Previous research has shown that perceived usefulness is a strong determinant of intentions to use, while ease of use is not a significant predictor [4]. Higher scores represent higher perceived usefulness and ease of use.

2.2 Materials and Equipment

The usability evaluations took place at the Center for Usability in Design and Accessibility. Participants completed user testing in a room separated from the researcher using a one-way window. Participants first completed an informed consent form then completed the tasks on a Dell desktop computer running on the Windows 10 operating system. Tasks that required the use of SPSS were completed using SPSS Version 23. IS was accessed in a Google Chrome browser. Microsoft Word was used by participants to type the answers for tasks. The participants' screen was recorded using Morae, and Google Hangouts was used for communication between the researcher and participants during the study. After each task, the participants were given surveys to measure perceived usability, ease of use, and usefulness. A post-test questionnaire was also given to participants regarding their preference of software.

2.3 Usability Testing

Tasks. Two tasks were developed. Each task consisted of a scenario to provide participants context for the current study. In the scenario, participants were told that they were students in a statistics course who needed to complete homework. The tasks were designed to simulate problems that students would typically encounter in introductory and intermediate statistics courses in the behavioral sciences.

The study used a within-subjects design and measured the performance of both software products for each participant. In addition, the software products and order of tasks were counterbalanced to account for order effects like learning and fatigue. Participants also tended to be somewhat experienced using SPSS, but had no previous experience with IS. Directions for both software were provided with the tasks to reduce performance bias associated with different levels of experience with the two software products.

The two tasks each contained different datasets. The first task involved a researcher interested in Body Mass Index who wanted to evaluate the effectiveness of an exercise program. The second task involved a researcher interested in standardized test scores and how they relate to overall school GPA. Each task was then broken down into six subtasks. The subtasks involved (1) data entry, (2) data cleaning, (3) data visualization, (4) descriptive analysis, (5) independent t-test, and (6) linear regression. Answers to the subtasks were typed in a Microsoft Word document.

Participants. Participants were 12 self-selected students ($N=12$) who responded to recruitment flyers posted around the psychology building at California State University, Long Beach, and were compensated 15 dollars (\$15 USD). There were five male and seven female participants. Five participants were pursuing an undergraduate degree in psychology. Seven participants were graduate students in the psychology department. The participants' mean overall GPA in statistics courses was 3.47 ($SD=0.39$) out of a four-point scale. Undergraduate students had the same mean GPA ($M=3.47$, $SD=0.37$) as graduate students ($SD=0.46$). To complete the tasks for both software products, participants needed previous experience with data analysis and writing APA formatted results. A prerequisite for participants in this study was the completion of intermediate statistics at

California State University, Long Beach. Information about participants' perceived level of understanding of statistics was obtained using a multiple-choice question with the possible answers of beginner, intermediate, or expert. Most users ($n = 9$) reported having an intermediate level of understanding of statistics.

Participants' experience and comfort with statistical analyses and writing results were measured using scales that ranged from one to five. From very inexperienced (1) to very experienced (5), participants indicated that they were somewhat experienced with performing data analyses ($M = 3.50$, $SD = 0.80$). From very uncomfortable (1) to very comfortable (5), participants indicated that they were between neutral and somewhat comfortable with writing statistical results ($M = 3.42$, $SD = 1.08$). Experience with statistical analyses and comfort with writing statistical results were strongly correlated, $r(10) = .89$, $p < .001$. Regarding experience with statistical software, participants had no previous experience using IS. Participants reported being somewhat experienced using SPSS ($M = 3.67$, $SD = 0.78$) and somewhat comfortable using SPSS ($M = 3.50$, $SD = 1.00$).

Procedure. Participants were first given the general background of the study and statistical software. Participants were then given time to explore the software prior to starting the task. For IS, participants were shown a brief video overview then given two minutes to explore the software. For SPSS, participants were given two minutes to explore the software prior to the task. Participants were not shown a video overview for SPSS as all participants had previous experience with the software.

After exploration of the software, participants began completing the tasks. Answers to the subtasks were entered in a Microsoft Word document. After finishing the first task, participants completed the SUS along with the surveys for perceived usefulness and ease of use. The same procedure applied for the second task. Participants were compensated after the questionnaires for the second task had been completed. The entire study lasted approximately 90 min.

3 Results

3.1 Analyses

Statistical analyses for the usability evaluation were conducted using SPSS Version 25. Task accuracy and time on task were analyzed using mixed design analyses of variance (ANOVA). Assumptions were inspected and violations were not found. The assumption of sphericity was not evaluated as there were only two levels for within-subjects variables. The results showed that completion rate for the regression subtask was lower than 70% for both IS and SPSS. Therefore, data for the regression subtask was not included in the analyses. Performance measures for effectiveness include task completion rate and task accuracy. Time on task was used as the performance measure for efficiency.

3.2 Task Completion Rate

All participants ($N = 12$) finished the first five subtasks for both software. For the linear regression subtask, eight participants completed the problem using SPSS ($n = 8$) and six participants completed the problem using IS ($n = 6$) (Fig. 1). Most participants did not finish the tasks due to a time constraint with the length of test session. The task completion rate was 67% and 50% for SPSS and IS, respectively. Consequently, data for the linear regression subtask was not used in the analysis on task accuracy and time on task.

Percent of Successful Completions by Subtasks ($n = 12$)

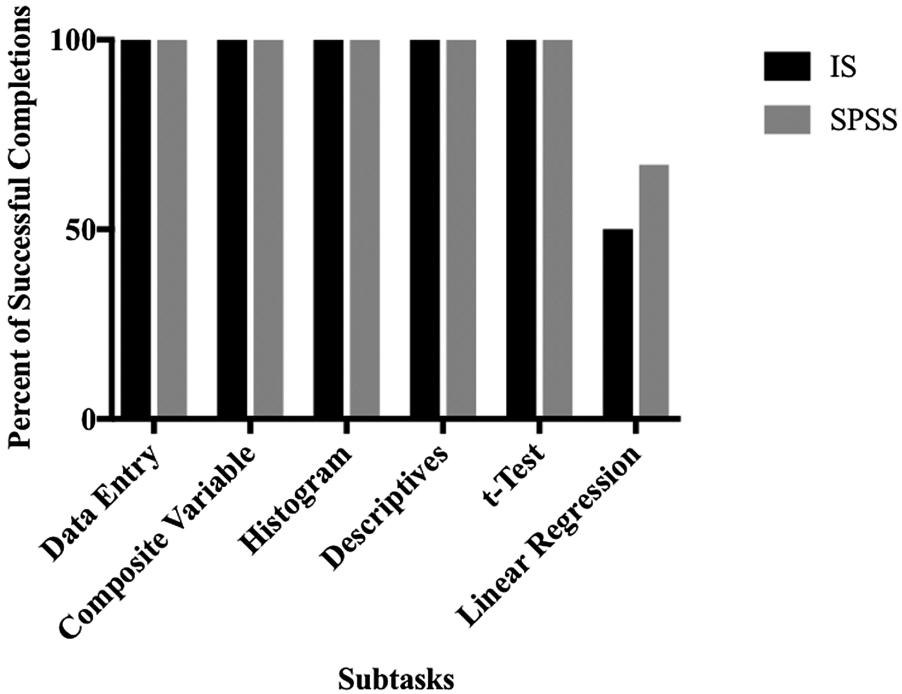


Fig. 1. Percent of successful completions for different subtasks and software.

3.3 Task Accuracy

Results indicate there was no significant difference in accuracy scores based on type of software used, $F(1, 8) = 0.38, p = .557$ (Fig. 2). The mean overall accuracy scores were 77.17 ($SD = 6.53$) and 78.83 ($SD = 7.91$) for SPSS and IS, respectively. There was also no significant difference in accuracy scores between the two tasks, $F(1, 8) = 0.02, p = .901$, or order of software used, $F(1, 8) = 0.22, p = .651$. The mean accuracy scores were 78.16 ($SD = 10.35$) and 75.91 ($SD = 6.62$) for Task 1 and Task 2, respectively.

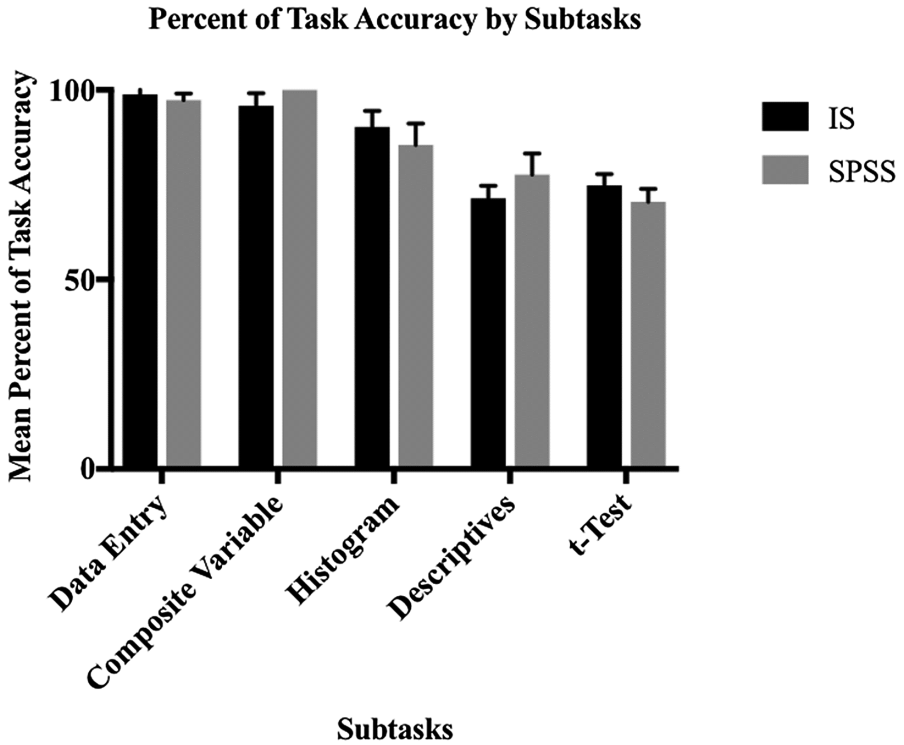


Fig. 2. Mean percent of task accuracy for different subtasks and software.

3.4 Originality Scores

Originality scores for answers using IS were not significantly different to originality scores for answers using SPSS, $t(11) = 1.99, p = .072$ (Fig. 3). However, the average percent of matching content to other sources was higher for IS ($M = 32.08, SD = 36.23$) compared to SPSS ($M = 8.17, SD = 22.06$). There were three answers for IS and one answer for SPSS that exceeded 70% matching content.

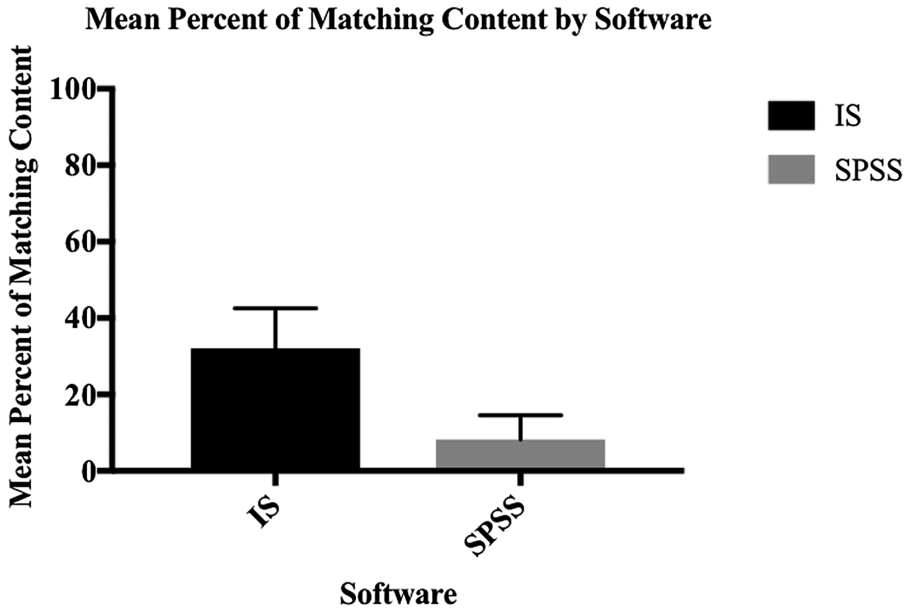


Fig. 3. Mean percent of content in participants’ answers that matched with other sources.

3.5 Time on Task

Results for time on task (in seconds) indicate there were no significant differences between IS ($M = 1697.58, SD = 1592.08$) and SPSS ($M = 1592.08, SD = 239.14$), $F(1, 8) = 1.59, p = .243$ (Fig. 4). There was also no significant difference in time on task between Task 1 and Task 2, $F(1, 8) = 0.15, p = .707$. However, a significant difference was found in mean completion time for order of software used, $F(1,8) = 8.16, p = .021, \eta_p^2 = 0.51$. Participants who used IS first spent a significantly longer time on tasks ($M = 1823.58, SD = 355.34$) compared to participants who used SPSS first ($M = 1466.08, SD = 192.10$).

There was also an interaction for time on task between type of software used and order of software used, $F(1,8) = 10.04, p = .013, \eta_p^2 = 0.56$. Participants who used IS for their first task spent a shorter time on the second task ($M = 1638.17, SD = 251.12$) compared to the first task ($M = 2009.00, SD = 363.62$). Similarly, participants using SPSS for their first task also spent a shorter amount of time on their second task ($M = 1386.17, SD = 90.56$) compared to the first task ($M = 1546.00, SD = 240.11$). However, the difference in mean time on task between the two software was significantly larger for participants who used IS first compared to those who used SPSS first.

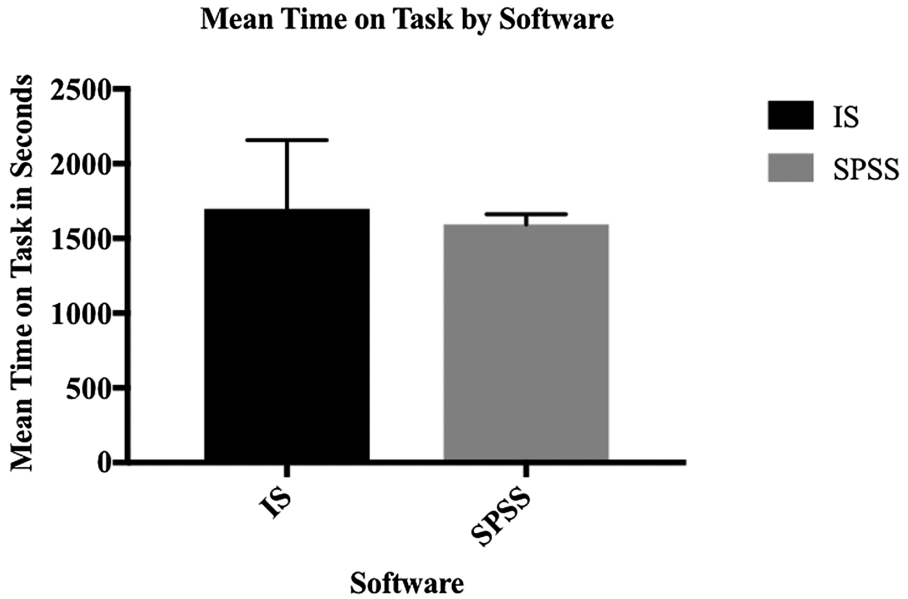


Fig. 4. Mean time on task in seconds for IS and SPSS.

3.6 Perceived Usability – SUS

SUS scores indicated that participants ($N = 12$) felt IS was significantly more usable compared to SPSS, $t(11) = 5.32, p < .001$. SUS scores for IS ($M = 83.33, SD = 11.74$) is considered to be at an acceptable level of usability [1]. SUS scores for SPSS ($M = 47.50, SD = 15.67$) were significantly lower and is considered to be at an unacceptable, but close to a marginal, level of usability (Fig. 5).

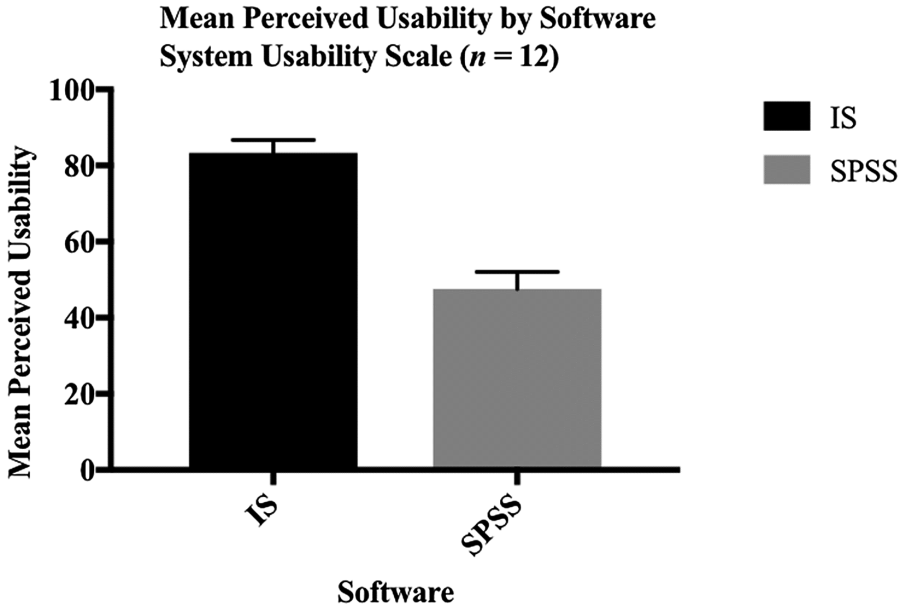


Fig. 5. Mean perceived usability scores for IS and SPSS.

3.7 Perceived Ease of Use and Usefulness – TAM 2

Results showed that participants felt that IS ($M = 5.88, SD = 0.84$) was significantly easier to use compared to SPSS ($M = 3.71, SD = 0.88$), $t(11) = 5.38, p < .001$ (Fig. 6). However, it was indicated that participants did not feel that IS ($M = 5.60, SD = 0.96$) was more useful than SPSS ($M = 5.06, SD = 1.02$), $t(11) = 1.34, p = .206$ (Fig. 7). Perceived ease of use for SPSS was not correlated with perceived usefulness, $r(10) = .16, p = .61$. The association between perceived ease of use and perceived usefulness for IS was also not significant, $r(10) = .57, p = .055$.

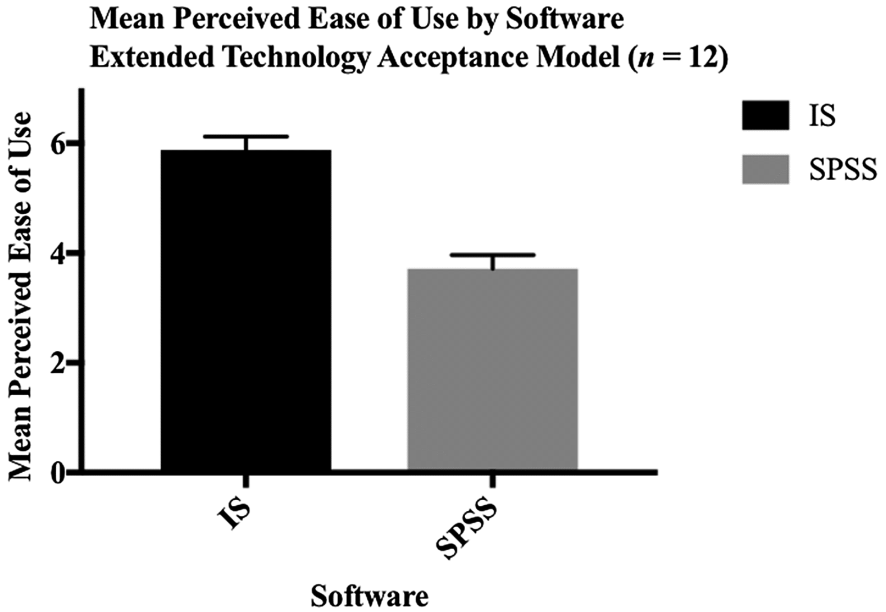


Fig. 6. Mean perceived ease of use rating for IS and SPSS.

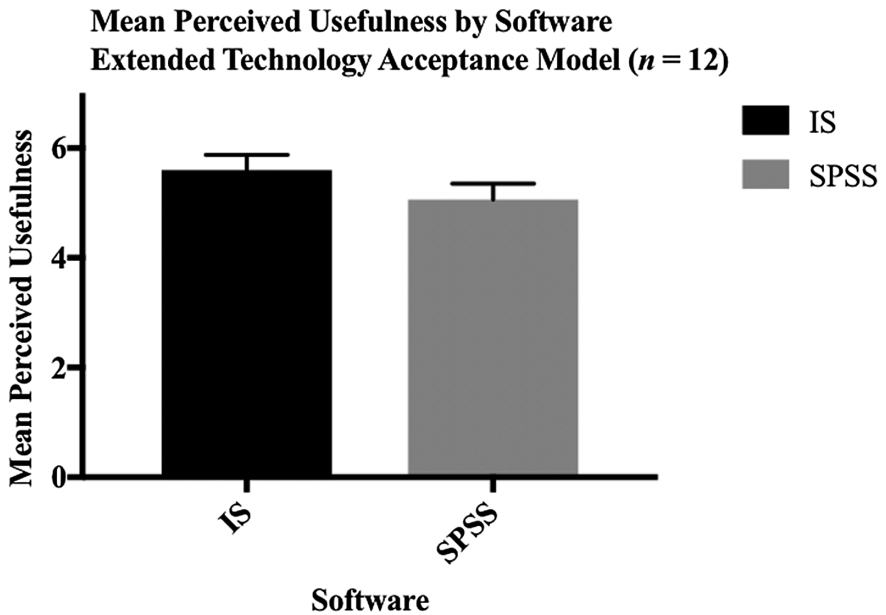


Fig. 7. Mean perceived usefulness rating for IS and SPSS.

3.8 Fear and Confidence Towards Conducting Statistical Analyses

Participants completed a post-test questionnaire after completing both tasks. Preference of statistical analysis software was measured using a scale with one indicating IS and seven indicating SPSS. Participants tended to prefer IS over SPSS ($M = 2.92, SD = 1.31$) (Fig. 8). Participants were also asked to rate if they felt that the statistical software reduced their overall fear of statistics using a Likert scale. A score of one indicated strongly disagree and a score of seven indicated strongly agree. Participants felt near neutral ($M = 4.75, SD = 1.28$) for IS, while disagreeing with the statement for SPSS ($M = 2.83, SD = 1.11$). Using the same scale, participants were asked to rate their agreement with a statement asking if using the statistical software increased their confidence in conducting statistical analyses. Participants agreed with the statement for IS ($M = 5.83, SD = 0.72$), but felt near neutral for SPSS ($M = 3.33, SD = 1.44$).

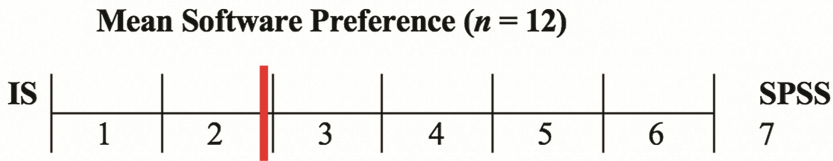


Fig. 8. Mean software preference of participants for IS and SPSS.

4 Discussion

The hypothesis of the current study was that IS would be more effective, efficient, and have higher usability compared to SPSS. To examine effectiveness, task accuracy was analyzed and there were no significant differences found. Overall, accuracy scores for IS and SPSS were fairly similar. One of the factors that led to higher effectiveness for SPSS when interpreting descriptive statistics is the amount of information provided in the output. While IS only provided the minimum and maximum values in the output, SPSS also provided the range. Many participants, despite having access to a calculator, did not calculate the difference between maximum and minimum values to obtain the range when using IS.

Another factor that influenced effectiveness of the software was the output presentation. The difference in accuracy scores for the t-test was largely due to poor performance in interpreting assumptions and including necessary information in the written results. The narrative format of IS output allowed users to easily obtain assumptions and check the format of their written results. SPSS output forced users to recall previous knowledge on assumptions and formatting requirements of APA results.

Scores for originality supported the conclusion that participants relied on the narrative format of IS output as a template when writing results. Though not statistically significant, IS had a higher mean percent of matching content compared to SPSS. Three participants using IS had written results with over 70% of their content matching with other online sources. Several participants also made a statement about wanting to copy

and paste the IS output to their written answers. This indicates potential issues, such as plagiarism, for students using IS for coursework. Past research also suggests that elaborative processing, though requiring more attention and time, results in improved learning and memory compared to shallow processing [6]. Despite having an interface that is more difficult to use, the need to interpret SPSS output in a contextual manner may lead to more effective student learning compared to IS.

To examine efficiency, time on task was inspected and a main effect for the order of software used was found. There was also an interaction between type and order of software used. A potential reason for this effect was that participants, who were inexperienced with IS, took a longer time to read and locate relevant information in the narrative output in the first task. However, if the narrative output was presented second, participants would likely skim the output due to fatigue or time constraints with the test session.

Perceived usability and ease of use scores for IS were significantly higher compared to SPSS. IS has an interface that is easier for users with a limited background in statistics to understand. The interface is more simplistic than SPSS and less overwhelming to use. The left menu only contains four major tabs and the menu options uses terminology familiar to users. This finding was expected as IS was designed for and marketed to users with limited statistical knowledge.

On the other hand, perceived usefulness was not significantly different for IS and SPSS. It is important to note that perceived usefulness is a stronger predictor of actual system usage than perceived ease of use [4]. While IS is simpler and easier to use, SPSS has more functionality. Many of the participants were seniors and graduate students that were taking or had taken advanced statistics courses. Some features in SPSS commonly used by intermediate or expert level users, including syntax and recently recalled dialogs, are not available in IS and may have influenced perceived usefulness ratings.

Overall, the post-test questionnaire suggested that participants tended to prefer IS over SPSS. Participants also felt that IS increased their confidence in conducting statistical analyses. This is most likely due to the easy-to-use interface combined with a narrative output. Past research has investigated the influence of self-efficacy, anxiety, and self-confidence on mathematics achievement in school [7]. The results suggested that confidence is strongest non-cognitive predictor of academic achievement. Thus, it is possible that IS may be an effective supplemental teaching tool for students learning statistics. The research hypothesis that IS would be significantly more effective, efficient, and have higher usability was only partially supported. Results did not support a significant difference in effectiveness and efficiency between the two software; however, IS was found to have significantly higher perceived usability and ease of use.

4.1 Limitations

The tasks and scoring rubric used in the present study were constructed by researchers with instructional experience in statistics. However, the tasks and rubric should be constructed in conjunction with a statistics subject matter expert to improve ecological validity. The 90-min time limit of the study was problematic as users were not able to complete all the tasks, which resulted in missing data for the regression subtask. Most participants were also showing signs of fatigue after the first task. Similar studies in the

future should be designed with shorter tasks to better account for participant fatigue. Participants for future studies should also have experience or training using IS. Ensuring that participants have comparable experience using both software would allow for the most accurate comparisons.

References

1. Albert, W., Tullis, T.: *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Newnes, London (2013)
2. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. *Int. J. Hum.-Comput. Inter.* **24**(6), 574–594 (2008)
3. Gkatzia, D.: Content selection in data-to-text systems: a survey. arXiv preprint [arXiv:1610.08375](https://arxiv.org/abs/1610.08375) (2016)
4. Guritno, S., Siringoringo, H.: Perceived usefulness, ease of use, and attitude towards online shopping usefulness towards online airlines ticket purchase. *Procedia-Soc. Behav. Sci.* **81**, 212–216 (2013)
5. McGrath, A.L.: Content, affective, and behavioral challenges to learning: students' experiences learning statistics. *Int. J. Scholarsh. Teach. Learn.* **8**(2), 6 (2014)
6. Schmeck, R.R.: Improving learning by improving thinking. *Educ. Leadersh.* **38**(5), 384–385 (1981)
7. Stankov, L., Morony, S., Lee, Y.P.: Confidence: the best non-cognitive predictor of academic achievement? *Educ. Psychol.* **34**(1), 9–28 (2014)
8. Venkatesh, V., Davis, F.D.: A theoretical extension of the technology acceptance model: four longitudinal field studies. *Manag. Sci.* **46**(2), 186–204 (2000)
9. Understanding the 2014–24 projections: Career Outlook. <https://www.bls.gov/careeroutlook/2015/article/projections-methodology.htm>. Accessed 27 Nov 2017
10. Wasserstein, R.: Communicating the power and impact of our profession: a heads up for the next Executive Directors of the ASA. *Am. Stat.* **69**(2), 96–99 (2015)



Comparative Study of Laptops and Touch-Screen PCs for Searching on the Web

Nicolas Debue^{1,2(✉)}, Cécile van de Leemput¹, Anish Pradhan³, and Robert Atkinson³

¹ Université libre de Bruxelles, Brussels, Belgium
ndebue@ulb.ac.be

² National Fund for Scientific Research, Brussels, Belgium

³ ANGLE Lab, Arizona State University, Tempe, AZ, USA

Abstract. This study compares the use of a laptop versus a touch-screen PC to perform web-based information search tasks. Thirty-six participants took part in a lab-based experiment. They were asked to use either a laptop or a touch-screen PC to seek information on the web and retrieve relevant pieces of information while their sessions were recorded. Cognitive load was measured through eye-related data and cortical activity (EEG) along with a self-reported scale. Main results indicated that participants who used the laptop outperformed those who used the touch-screen PC, with more relevant webpages bookmarked ($F = 9.678$, $p = .004$) and more relevant elements retrieved ($F = 6.302$, $p = .018$). Participants with the touch-screen PC also spent more time on each webpage than their counterparts ($F = 9.2141$, $p = .005$). These results suggest that using the touch-screen PC required more mental effort, which is supported by cognitive load measurements. Linear mixed-model analyses showed significant differences across devices in both pupil size variation ($F = 3.692$, $p = .05$) and EEG-based cognitive load index ($F = 5.181$, $p = .028$). This study raises issues about whether touch-screen computers are suited for every computing needs.

Keywords: Information search · Cognitive load · Performance · Mobile device

1 Introduction

Since the release of the first iPad in 2010, billions of tablet computers have been sold all over the world¹. From playing games at home to presenting slideshows at work, tablet computers are now used for a variety of purposes and in a wide range of contexts [1]. Although many had predicted that tablet PCs sales would overtake the traditional PCs market by 2018², reality has not met expectations with a steady decrease in tablet PC sales since 2015³. According to a recent Gartner's report⁴, mobile device adoption in the workplace has been lower than expected with workers still massively relying on desktop computer or laptop for their computing needs. For workers, tablet PCs are still too limited

¹ <http://www.emarketer.com/Article/Tablet-Users-Surpass-1-Billion-Worldwide-2015/1011806>.

² <https://www.gartner.com/doc/2945917/forecast-pcs-ultramobiles-mobile-phones>.

³ <https://www.idc.com/getdoc.jsp?containerId=prUS42272117>.

⁴ <https://www.gartner.com/newsroom/id/3528217>.

when it comes to content creation such as writing and editing document, designing presentation or using spreadsheet⁵. To fight tablet PCs decline, manufacturers have started marketing new devices such as “detachables” and “convertibles” that combine the portability and the touchscreen of a tablet PC with the power of a laptop. A detachable tablet is a device that is composed of a touchscreen and a detachable keyboard (such as Microsoft Surface Pro) while a convertible is a notebook with a touchscreen that can be flipped or folded depending on the model (Lenovo IdeaPad...). With those new devices, users are given the choice between several interaction modes as they can use the regular mouse/keyboard input, the touchscreen only, or a combination of both. However, despite the ever-increasing penetration of these new devices in firms⁶, only few studies have examined whether the interaction mode can impact user’s performance.

Ostrowski [2] have shown that touch-based interactions could help to reduce the overall load on mental resources compared to mouse and keyboard, thus improving performance on the touch-screen devices. Wang [3] claimed that using gestures mimics common real world gestures and avoids the split attention effect due to the use of the mouse and keyboard. Nevertheless, other researchers did not find any learning gains when tablet PCs are used instead of traditional computers [2]. Since the emergence of touch-screen devices, many studies have compared physical keyboards and on-screen keyboards [4–6]. Varcholik *et al.* [6] found that using a virtual on-screen keyboard to encode text on a device decreases the speed and increases typing errors when compared to physical keyboards. Furthermore, participants reported being less satisfied with the on-screen keyboard. According to Findlater and Wobbrock [5], users devoted more attention to typing due to looking at the on-screen keyboard to reduce their error rate. Along with a touch-based interaction mode and an onscreen-keyboard, tablet PCs usually feature smaller displays than laptops with screen sizes ranging from 7 to 11 in. for a “slate” tablet and from 11 to 13 in. for most of the detachables and convertibles. This raises concerns for potential detrimental effects on user’s performance due to increase scrolling, which has been reported to lower text comprehension [7] and task efficiency [8].

In this paper, we investigate the impact of the interaction mode on information search performance, since it has been reported to be one of the most frequently performed activities on IT devices at work [9, 10]. Browsing the web is also a commonly performed activities on tablet PC with around 50% of users who engage in such activity on their touch-screen devices [11]. This paper aims to tackle some of the ergonomic issues raised by the increasing use of touch-screen computers at work by exploring the impact of the interaction modes on users’ behavior while they are seeking information on the internet. In the next section, we will briefly introduce our main theoretical framework and our hypotheses will be outlined. Then, we will describe the experiment conducted and the related method and material. Results will be presented and discussed thereafter.

⁵ <https://www.spiceworks.com/marketing/resources/reports/rise-mobile-empire/>.

⁶ <https://www.idc.com/getdoc.jsp?containerId=prUS42332117>.

2 Theoretical Background

2.1 Information Seeking Models

Many models have been developed to describe how people search information [12–15]. Marchionini's model [16] describes information searching activity as an eight-step problem-solving process: recognize and accept the problem, define and understand the problem, select the source, formulate the query, execute the query, examine the results, extract information and reflect, iterate or stop the research. While Marchionini was the first to emphasize the iterative nature of the search process in digital environments, his model lacked a clear description of how people interact with the systems and overlooked the role of cognitive and psychological factors [17]. Building on Marchionini's model, Sharit and colleagues [18] developed a model specifically related to search engines that emphasizes the role of working memory and web-related knowledge on information-seeking (IS) performance. Their model describes IS as a multi-stage process "whereby the problem-solver's knowledge and other mental representations are manipulated to achieve a goal" [18, p. 3]. First, the problem is identified and broken into several goals and subgoals. Then, search terms are generated and translated into queries in a search engine. Webpages matching the initial goals are visited and the information retrieved is then compared to these goals. The information-seeker stops the research when he/she judges that a sufficient amount of information has been found or when he/she gives up. This process is highly iterative as the mental representations can be modified at each step of the process, thereby leading to a reformulation of the problem and the corresponding goals.

This model also points out the role of web-related knowledge that can impact the formulation of the queries as well as the relevance judgment of the webpages retrieved. Furthermore, the authors emphasize the importance of working memory in keeping the user oriented throughout the IS process, especially as the difficulty of the problem-solving task increases. While searching on the web, the user has to maintain the goals in working memory, set and apply a search strategy, as well as process and store information. Both the amount of information that has to be maintained in working memory and the remaining resources available to perform the search can impact the performance in IS [18]. In line with the objectives of this study, it is worthwhile questioning whether seeking information using a specific device could impact the user's mental resources in terms of working memory and thereby, their performance in IS activity.

2.2 Cognitive Load

Cognitive theories constitute a relevant framework to explore how touch-screen computers can impact users' mental resources. Originally developed in the field of instructional design, Cognitive Load Theory [19] and her sister theory, the Cognitive Theory of Learning with Media [20], are based on the same assumptions. Firstly, that the human cognitive architecture is made up of two dependent structures: (a) a working memory that actively selects and integrates incoming information with prior knowledge and mental models; (b) a long-term memory that stores a potentially unlimited amount

of information in the form of schemata. Secondly that the working memory has limited resources, and activities that demand attention compete for these resources. Cognitive load refers to this limited capacity and is described as the mental cost of a specific task, for a particular individual in a given context. According to these theories, cognitive load is multifactorial and can be divided into three types of cognitive load: intrinsic, extraneous and germane loads. Intrinsic load is related to the material to be learnt (the interactivity of the elements) and the user's prior knowledge. It is usually related to the task demand and task difficulty. Extraneous load deals with the mental resources devoted to elements that are not directly related to the task at hand, and is often linked to the presentation format. Germane load is described as the mental resources required by schemata acquisition and automation in working memory. Since the working memory capacity is limited, an increase of extraneous load is correlated to a decrease of germane load and, consequently, results in lower learning outcomes or performance.

Different methods allow the measurement of cognitive load: subjective ratings, performance-based measures and psychophysiological measures (see [21] for a review of these methods). Subjective ratings assume that an individual is able to report how much mental effort has invested in the task undertaken. Researchers have used multi-dimensional scales to discriminate between the different load factors and distinguish intrinsic, extraneous and germane load [22]. Psychophysiological measures have several advantages over subjective ratings. Since they are based on bodily responses, they allow measurement at a high rate and with a high degree of sensitivity [23]. Moreover, they do not require an overt response by the subjects and, in this way, are considered as a direct and "objective" way to infer mental activity. In this paper, we focused on both types of measures, which will be detailed in the Method section.

Given the aforementioned elements, we hypothesize that browsing the web on a touch-screen computer is likely to increase webpage scrolling which will result in a higher demand upon cognitive resources and a rise in extraneous load. Moreover, keying errors are more frequent on onscreen keyboards than on their physical equivalents and therefore, require much more attentional resources.

H1. Using a tablet computer will generate a higher level of extraneous load compared to a laptop

As pointed out by Sharit and colleagues [18], information search on the web requires the employment of a substantial amount of mental resources. The seeker has to keep the goal of the search in working memory while making decisions about the search results' relevance until the task is successfully achieved. Assuming that using touch-screen computers requires more mental resources and that the working memory has a very limited capacity, we hypothesize that the use of a touch-screen computer hinders user performance in information search when compared to searching on a laptop.

H2. Laptop users will outperform touch-screen computers users in online search.

3 Method

3.1 Sample

Thirty-six students from Arizona State University took part in this experiment. There were 17 males and 19 females with a mean age of 21.36 years ($SD = 1.86$). They had normal or corrected-to-normal vision and did not report any attention disorder. Participants were given a compensation of 25 dollars for participating in this study and signed an IRB-approved consent form at the beginning of the study.

3.2 Control Variables

Numerous studies have indicated that previous knowledge or experience of the internet, experience of the search engines and of the devices used can impact the efficacy of the search process [24]. Familiarity with the devices was controlled for by asking the participants to report the number of hours spent on each device on a weekly basis (for both laptop and tablet PC). Device self-efficacy was measured via a scale adapted from Compeau and Higgins [25] and modified according to the device to be used. The scale was made of 6 items referring to the ability of the user to use the device (skill based) such as “I could complete a new task using a laptop/tablet PC”. Cronbach’s alpha were .765 for the laptop scale and .772 for the tablet scale respectively. Information search perceived self-efficacy was measured thanks to a 13-item scale (Cronbach’s alpha was .847) developed by Bronstein [26]. Items such as “I can usually find the information I need” assessed users’ perceived ability to perform information search. Participants were asked to rate the extent to which they agreed/disagreed with the statements on a 5-point Likert scale ranging from “Strongly disagree” to “Strongly agree”.

Working memory capacity has been indicated as related to performance in multi-tasking on the web [27] and information search tasks [28]. In this study, we used the working memory capacity test developed by Oswald et al. [29]. Due to time constraints, only operation span was evaluated for each participant. However, operation span has been specifically related to performance in IS tasks [30]. Sets of arithmetic operations were presented to participants who had to judge whether the problem was true or false (e.g. $25 + 5 = 30$). After each problem, a letter was presented and participants had to recall all letters in the right order after the set of equations. Set size ranged from 3 to 7 and there were three administrations for each set size. The test took 15 min on average. Independent samples t-tests performed on the control variables did not yield any significant difference across groups (touch-screen versus laptop) nor gender.

3.3 Measurement Tools, Information Search Tasks and Dependent Variables

Cognitive Load. Extraneous load represents the load related to the mental resources devoted to the elements irrelevant to the learning tasks [19]. In the usability context, extraneous load has been identified as the load generated by poor layout and usability issues [31] or disorientation in a hypermedia [32]. Based on the usability heuristics [33], extraneous load was measured with 4 items referring to: (1) the ease of navigation on

the Web (“Navigating between pages was a problem”); (2) the amount of information displayed (“the amount of information displayed on the screen was appropriate”); (3) the ease of interaction (“It was easy to interact with the device”); (4) the perceived disorientation (“I could identify easily on what page I was and where to go next”). As perceived extraneous load was measured after each information search task, the scores were averaged over the 5 tasks. Cronbach’s alpha coefficients were computed to estimate both the internal consistency of the scales and the reliability of the repeated measurements. All coefficients showed good to very good reliability.

Physiological measures were collected to obtain a more sensitive, reliable and unbiased measure of extraneous load. Techniques such as EEG, galvanic skin response, heart rate variability or fMRI have been used to assess physiological proxies of mental activities but eye-related data is still amongst the most widely explored. Eye-tracking is a non-intrusive and cost-effective method that allows one to observe the user’s attention allocation through the gaze position. Multiple indexes can be collected such as gaze point position, number and frequencies of eye blinks, duration of fixation and saccades (see [34, 35] for a comprehensive review of these measures). The pupillary response is one of the most popular measure of cognitive load due to the fact that the pupillary reflex is under the control of the autonomous nervous system and cannot voluntarily be controlled by the subject. Relationship between an increase of cognitive load and pupil diameter has been described in various contexts (varying from such as simple cognitive tasks [36, 37] to naval simulators [38]; driving [39]; e-learning [40]; e-shopping [41] and an AI web-based tool [42]).

As pointed out by Chen and colleagues, “there can be no single measure that can be recommended as the definitive measure of mental load” (p. 35) [43]. In an attempt to provide reliable measures of cognitive load, along with the eye-related data, cortical brain activity was also measured using an electroencephalograph (EEG). The use of EEG has been validated to measure mental workload and task engagement in various environments and tasks (see [44–46] for more details). Berka *et al.* [44] used quadratic discriminant functional analysis on the decontaminated EEG signal to create a task engagement index with four levels: high engagement, low engagement, distraction and sleep onset. A probability of cognitive state is then provided for each second. According to Berka *et al.* [45], EEG-engagement is related to visual scanning, sustained attention and information gathering. While some researchers have successfully used this index to show a decrease in cognitive load when using a specific software [47] others did not find any difference in engagement in subjects using video games. In order to find an index that is more suited to the context of web-based information search, we followed the procedure described by Sénécal *et al.* [48]. They calculated a cognitive load odds based on the EEG-engagement metric using the equation below:

$$\text{Cognitive Load Odds} = \frac{\text{Probability of high engagement} + \text{Probability of low engagement}}{\text{Probability of distraction} + \text{Probability of sleep onset}}$$

They successfully used this index to discriminate across variations of the user’s cognitive load when visiting or revisiting websites and the impact of website familiarity on cognitive load. Pupil size variation from the baseline and EEG – Cognitive Load Odds were averaged over the tasks in order to obtain a measure for the whole

experimental run. For the EEG-Cognitive Load Odds (EEG-CLO), a logarithmic function was applied before averaging over the different tasks. In this experiment, pupil size was considered as a proxy measure of overall load and EEG-CLO index was considered as a measure of extraneous load.

Information Search Tasks. The search tasks were designed following the simulated work situations principles [49] which means that the task describes the source of information need and the environment, in order to make clear the objectives of the search to the seeker. Effort has been made to generate information search (IS) tasks that were interesting for our population of interest and that motivated them to perform realistic searches. Based on previous research on IS tasks [30, 50], we defined two types of tasks: fact finding and information gathering. The fact finding tasks required the retrieval of one or more specific pieces of information while information gathering tasks were less defined and required the collection of several pieces of information on a given topic. There were three structures for FF tasks: Simple, Hierarchical and Parallel. For FF Simple tasks, one piece of information had to be retrieved to achieve the goal. Hierarchical concerned a deeper search as they had to retrieve a number of pieces of information about the same topic but located at different levels of depth. Conversely, Parallel search dealt with multiple concepts that exist at the same level of depth (breadth search). Table 1 summarizes the tasks and their characteristics.

Table 1. Instructions, type and structure of the tasks

| Task | Instruction | Type | Structure |
|-------|---|-----------------------|--------------|
| FF_S0 | Find out the date of the next supermoon eclipse visible from Phoenix | Fact finding | Simple |
| FF_S1 | Find out the GDP of China in 2015 | Fact finding | Simple |
| FF_H1 | Find out information about coral snakes, its color and dangerousness | Fact finding | Hierarchical |
| FF_P1 | Find out information about the acceptance rates in US colleges over the last 3 years and the next 3 | Fact finding | Parallel |
| IG_H1 | Collect information about artificial sweeteners consumed in the everyday life | Information gathering | Hierarchical |

Performance Metrics. Performance was defined as a second order concept that encompasses three dimensions: (1) search outcomes; (2) search effort and (3) depth of search. For each information search task, participants were asked to locate one or several pieces of information using the web-based search engine Google. Once the information was located, they had to clipboard the relevant piece(s) of information and bookmark the corresponding page(s). Search outcome was based on two metrics: the number of elements retrieved weighted by their relevancy and (2) the number of pages bookmarked weighted by their relevancy. All bookmarks and elements in the clipboard were judged as task-relevant. Search effort refers to the effort spent to achieve the goal of the task and is compounded of two elements: the time spent for each bookmark and the time per page. Finally, depth of search reflects the motivation of the participant to seek for

information. Since they were allowed an indefinite period to perform the search, one could say that a longer, more motivated search was more likely to end up in a more complete answer, more bookmarks and thus, better performance. Accordingly, we defined depth of search as the time spent on the task, the number of queries formulated and the number of webpages visited. Scores were added up across all tasks in order to compute an overall score for each device.

3.4 Apparatus

In the laptop condition, we used a Dell Latitude E6540 with an Intel Core i5 quad processor and 8 GB RAM. The monitor was a 15.4 in large screen with a resolution of 1920×1080 . No external mouse or keyboard was provided in order to make the participants use the integrated keyboard and touchpad. For the tablet PC condition, we used a Lenovo Yoga 13 tablet PC with an Intel Core i7 and 8 GB RAM. The monitor was 13.3 in large with a resolution of 1600×900 . The tablet PC was folded so that participants were only provided with the touch-screen to interact with the device. Both computers were running Windows 8 and used Google Chrome as their default internet browser.

EEG measurement involves detecting the fluctuation of voltage potential generated by large groups of neurons in the brain. The EEG signal was acquired using the B-Alert X10 device from Advanced Brain Monitoring. This device allows us to remotely acquire data of brain activity using a wireless set of nine electrodes (F3, F4, Fz, C3, C4, Cz, P3, P4 and POz) sampled at 256 Hz. B-Alert proprietary software uses an artifact decontamination algorithm to account for electrical interferences, eye blinks or motor movements. It computes two composite metrics: a cognitive workload index and a cognitive state index (see [44] for the technical details).

Eye tracking data was collected via a Tobii X2-60 remote eye tracker sampled at 60 Hz. For the laptop condition, the tracker was attached below the screen in order to track the eyes even when the eye-lids were partially closed. For the touch-screen condition, both the tablet PC and the tracker were fixed on a tailor-made mobile device stand that allowed the users to use the touchscreen without putting their hands/arms in front of the tracker. iMotions software version 5.7 was used to display the questionnaires and tasks and allowed the integration and synchronization of the EEG and eye tracking signals.

3.5 Protocol

The study was conducted in a testing room at Arizona State University. The participants were first escorted into the lab by the examiner, then they were seated and given an informed consent form. Once the form was viewed and signed, the study procedure was explained and the participant had to complete the computer-based working memory capacity test. Then, the EEG headset was placed onto the participant's head, an impedance check was run before starting the ABM calibration procedure. This calibration takes around 15 min and consisted of 3 computerized tasks in which visual (colorful shapes) and audio stimuli had to be identified.

Then, the eye tracking was calibrated and participants had to gaze at a blank screen for a 10-s baseline. Before moving to the information search tasks, they had to fill out an online questionnaire including questions about their age, gender and the control scales (cf. supra). Instructions regarding the information search tasks were provided and a video played showing a trial task and explaining what was expected to perform the task. The actual tasks were then displayed in a randomized order. For each task, they were asked to retrieve one or more pieces of information, clipboard the relevant items and bookmark the corresponding pages until they considered they had provided enough relevant elements to the problem presented. No time limit was set to complete the tasks. After each search task, they had to report their level of cognitive load for the task just performed. Once all tasks were completed, participants were thanked for their participation, compensated, and given information on obtaining the results of the study. The whole experimental run took around 2 h.

4 Results

All data transformations were performed on the open-source software R [51] and statistical analyses using the statistical suite SPSS. The assumption of normality was met for all the variables included in our analyses. Independent samples t-test were used to analyze the differences in scores across the two conditions. Table 2 provides the mean and standard deviations for the main performance metrics as well as the t-statistics, corresponding p-values and effect sizes.

Table 2. Means (SD) and results from the t-tests for the time on task, depth of search, efficacy and search effort

| | Laptop | Tablet | Df | <i>p</i> | Cohen's <i>d</i> |
|------------------------------------|-----------------|-----------------|----|----------|------------------|
| Mean search time (sec) | 284.03 (139.72) | 300.35 (101.53) | 34 | .705 | |
| Depth of search | | | | | |
| Number of webpages visited | 4.16 (1.65) | 2.74 (1.09) | 34 | .005 | 0.96 |
| Number of queries | 4.67 (1.65) | 3.14 (1.26) | 34 | .004 | 1.04 |
| Efficacy | | | | | |
| Number of items in the clipboard | 3.46 (1.78) | 1.43 (0.64) | 34 | <.001 | 1.51 |
| Number of bookmarks | 9.29 (3.01) | 6.88 (2.61) | 34 | .018 | 1.30 |
| Search effort | | | | | |
| Mean time per bookmark (sec) | 32.99 (19.1) | 51.32 (23.45) | 34 | .021 | 1.10 |
| Mean time spent on each page (sec) | 73.08 (36.4) | 122.35 (37.96) | 34 | <.001 | 1.32 |
| Perceived extraneous load | 2.55 (0.77) | 3.32 (.79) | 34 | .008 | 0.96 |

Results showed that, when averaging time across all tasks, there was no statistically significant difference in time between the two devices. Regarding performance in information search tasks, our results indicated that the depth of search and efficacy were lower for those who used a touch-screen PC compared to those who used a laptop. Fewer

webpages ($M = 2.74$) were visited and fewer queries were formulated ($M = 3.14$) on the touch-screen condition compared to the laptop condition. The t-test analysis showing significant differences for both the number of webpages visited ($p = .005$, $d = 0.96$) and the number of queries generated ($p = .004$, $d = 1.04$). Similarly, the participants who used the touch-screen PC obtained a worse overall efficacy with almost half the number of elements copied into the clipboard ($M = 1.43$) than in the laptop condition ($M = 3.46$). Those on the touch-screen PC bookmarked 25% less webpages ($M = 6.88$) than those on the laptop ($M = 9.29$). Regarding the search effort, the results indicated that touch-screen group spent significantly more time for each web page bookmarked than their counterparts on the laptop ($p = .021$, $d = 1.10$). A significant difference was also found when looking at the average time spent per task. As showed in Table 2, 122.35 s were spent on each webpage visited on a touch-screen PC against only 73.08 s when visited on a laptop ($p < .001$, $d = 1.32$).

We hypothesized that these differences in performance might be related to variations of cognitive load. Table 2 shows that using the touch-screen PC led to a higher level of perceived extraneous load ($M = 3.32$) compared to a laptop ($M = 2.55$), this difference being statistically significant ($p = .008$, $d = 0.96$).

Along with this self-reported measure, cognitive load was measured by two physiological metrics: pupil size variation from the baseline and EEG cognitive load (EEG-CLO). Linear Mixed Model analyses were performed using the MIXED procedure in SPSS (see [52] for an introduction to those models). The device used was defined as fixed factors. To account for within-subject variability related to physiological data and the non-independence of repeated measures, a random intercept was defined along with the fixed effect. Relationships amongst the residuals were directly estimated in the model using a Variance Component covariance matrix. Intra Class Correlation coefficients were computed by dividing the unexplained variance of the residuals by the variance of the random factor, which gives the percentage of variance explained by the random effect.

Table 3 shows a decrease in pupil diameter for those who used the tablet PC compared to those who searched on the laptop. Regarding EEG data, results showed a significant increase of EEG-CLO for the touch-screen condition ($M = 1.87$) than for the laptop condition ($M = 1.04$); $F = 5.27$, $p = .028$. Finally, ICC coefficients indicated that for the two measures, the random intercept factor explained the majority of the residual variance with coefficients of 74.9% and 87.7%. Pearson’s product-moment correlation analyses were run to assess the relationships amongst loads but no significant correlations were found.

Table 3. Means (SD) and linear mixed model analyses of the pupil size variation and the EEG-based index of cognitive load

| | Laptop | Tablet | F | <i>p</i> | AIC | ICC |
|---|-------------|--------------|------|----------|--------|------|
| Pupil size variation from the baseline (mm) | 0.007 (.01) | -0.037 (.01) | 4.10 | .05 | 348.24 | 74.9 |
| EEG-CLO (system unit) | 1.04 (0.64) | 1.87 (1.34) | 5.27 | .028 | 209.95 | 87.7 |

Notes. N = 33; AIC: Akaike’s Information Criterion; ICC: Intra Class Coefficient

5 Discussion

In an effort to explore the impact of the use of touch-screen PCs, we investigated whether seeking information on the web differs when performed on a touch-screen computer versus on a laptop. A lab-based study was carried out to shed light on this issue as well as to explore the underlying mechanisms that could explain the impact of a specific device on information search (IS) performance. We hypothesized that using a tablet PC would decrease performance in IS compared to a laptop, since a touch-screen PC requires more mental resources to be used, thereby reducing those available to perform the search.

The results of this study provide evidence that using a touch-screen computer to seek information on the web results in a drop in performance in IS (H2 is supported). Fewer queries were formulated, fewer webpages were visited and therefore, participants retrieved fewer relevant elements to achieve the goals of the tasks. They either did not find any relevant element or they provided a less complete answer than those who had sought information on a laptop. It must be noted that all elements saved into the clipboard and all pages bookmarked were task-relevant, regardless of the device used.

While fewer webpages were visited, the touch-screen PC group spent much more time on each webpage which could indicate that reading required more effort. Consequently, those users might have engaged in a more in-depth exploration of the webpage, trying not to jump from one page to another too quickly so as to minimize the mental effort in navigating the web. Also, of interest is that one would expect that the search should have taken longer as a result of the strain generated by the device, yet our results showed no difference on the average time spent per task. A possible explanation for this might be that participants interrupted the search process earlier because they were frustrated or disoriented in the task. As described by Sharit's model [18], the ability to stay oriented is a critical determinant of the success of the search process.

With regard the Cognitive Load Theory framework [19], we suggested that this drop in performance might be caused by an increase of extraneous load. Our findings are in line with this claim since participants who used the touch-screen PC reported a higher level of perceived extraneous load (H1 is supported) than those on the laptop. It means that they experienced difficulties with navigating between pages and interacting with the device and they judged that the amount of information displayed on the screen was not optimal. It can therefore be assumed that the smaller screen, the on-screen keyboard and the touch-based interaction mode specific to those type of devices may impose a burden on the user's mental resources. Interestingly, we obtained these results despite using a medium-sized touch-screen PC (13.3") although some devices feature much smaller screens (below 10"). Using such devices would have certainly resulted in more marked differences.

Along with these self-reported measures, we gathered physiological data in order to get more sensitive, reliable and unbiased measures of cognitive load. While there is still much debate with respect to the multimodal measures of cognitive load [43], our findings provide some interesting food for thought. First, it is noteworthy that our proxy measures were not correlated together which indicates that there are not likely to be related to the same underlying factors. As expected, a negative pupil size variation was found in the touch-screen condition. Given that pupil size variation has been related to information

processing [34] in working memory, this could account for the decrease in both depth of search and efficacy. Our EEG-based measure of cognitive load (EEG-CLO) was higher in the touch-screen condition, which supports our assumption that using to search information on the internet leads to a higher level of extraneous load. However, there is still a lack of understanding about the specific elements that generate this burden. It is possible that averaging the EEG-CLO index over the whole experimental run has hidden subtle relevant variations of the mental load during the tasks. For instance, previously an EEG study has shown differences in mental workload when people select links or read text online [53], events that occur and must be analyzed at a higher temporal resolution. Further investigation should be undertaken to precisely identify what are the main elements users are struggling with when searching the web on a touch-screen device.

5.1 Limits

The present research has several limitations that should be noted. First, as with most lab-based studies, our sample size was relatively small (due to the duration of the experimental procedure). Our results showed, however, that despite this limited sample size, our statistical analyses mostly yielded small p-values and meaningful effect sizes. Second, our findings should be cautiously interpreted when generalizing to other populations since our participants were only young adults. Older adults and elderly may be less familiar with touch-screen devices and may be more disoriented while using such devices [54]. Third, a folded convertible laptop was used instead of a slate tablet PC. Although participants had to use gestures to interact with this device, they were not allowed to handle it due to the use of the eye tracking system. One might argue this may have hampered the interaction and worsened performance. On the other hand, the similarity of the devices (same operating system, same browser) limited the number of potential confounding variables. Furthermore, since the size of the convertible laptop screen was greater than a standard slate tablet PC, we believe that it could have “softened” our results by reducing the need for scrolling. Replicating this study on an off-the-shelf slate tablet PC should thus be considered. Finally, the study was conducted in a controlled environment and participants were asked to perform only one information search at a time. In the workplace, it is most likely that workers are using their devices on the go while doing several tasks in parallel. As multitasking has a strong negative impact on memory encoding [55], it raises the possibility that the impact of using a tablet could be more pronounced in such conditions.

5.2 Future Research

To develop a full picture of the impact of touch-screen devices in the workplace, additional studies will be needed that extend the nature of the tasks involved, as well as taking into account the effect of multitasking on users’ mental resources. It would be interesting to extend the test scenario by including tasks such as checking and writing emails or word processing, in order to best encompass the variety of the tasks performed on IT devices at work. There is also much room for further progress in refining our understanding of the objective measurement of cognitive load. Future research should

seek to extend our understanding of the relationships between physiological measures and the underlying cognitive processes in the context of web browsing. Another important avenue for further research concerns the effect of motivational factors on users' behavior. Techniques such as galvanic skin response or emotion recognition software could be considered to obtain objective measures of users' emotions or cognitive engagement while using different devices.

5.3 Conclusion

The present research has demonstrated that the device used to search online can impact users' performance. Regardless of this, it is undeniable that touch-screen computers such as slate tablet PC, detachables and convertibles have the potential to be of great benefit in the workplace, thanks to their portability, usability and interactivity. However, they might not be suited to every activity required in the workplace and special care needs to be taken when introducing new technology into the firms.

Acknowledgments. This research was supported by the Belgian National Fund for Scientific Research. We thank our colleagues from the Advanced Next Generation Learning Environments at Arizona State University, particularly Maria Elena (Helen) Chavez-Echeagaray who provided insight and expertise that greatly assisted the research. We would also like to express our gratitude to iMotions Inc. that provided us a trial version of iMotions for conducting this study.

References

1. Müller, H., Gove, J., Webb, J.: Understanding tablet use: a multi-method exploration. In: Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services, pp. 1–10. ACM, New York (2012)
2. Ostrowski, C.: Touching our way to better conversations: how tablets impact cognitive load and collaborative learning discourses. In: Simonson, M. (ed.) Proceedings of the Annual Convention of the Association for Educational Communications and Technology, pp. 159–167, Jacksonville, FL (2014)
3. Wang, T., 王天宠: The effect of tangible user interface on iPads in learning behavior: a case study of international schools in Hong Kong (2012). http://dx.doi.org/10.5353/th_b4853944
4. Chaparo, B., Nguyen, B., Phan, M., Smith, A., Teves, J.: Keyboard performance: iPad versus Netbook. *Usability News* **12**, 1–9 (2010)
5. Findlater, L., Wobbrock, J.O.: From plastic to pixels search of touch-typing touchscreen keyboards. *Interact.* **19**, 44–49 (2012)
6. Varcholik, P., LaViola, J., Hughes, C.: Establishing a baseline for text entry for a multi-touch virtual keyboard. *Int. J. Hum.-Comput. Stud.* **70**, 657–672 (2012)
7. Sanchez, C.A., Wiley, J.: To scroll or not to scroll: scrolling, working memory capacity, and comprehending complex texts. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **51**, 730–738 (2009)
8. Botella, F., Moreno, J.P., Peñalver, A.: How efficient can be a user with a tablet versus a smartphone? In: Proceedings of the 15th International Conference on Human Computer Interaction, pp. 64:1–64:9. ACM, New York (2014)
9. Forrester Consulting: 2013 Mobile Workforce Adoption Trends (2013)

10. Google: Consumer Barometer from Google. <https://www.consumerbarometer.com/en/graph-builder/?question=M7b2&filter=country:belgium>
11. Müller, H., Gove, J.L., Webb, J.S., Cheang, A.: Understanding and comparing smartphone and tablet use: insights from a large-scale diary study. Presented at the Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction (2015)
12. Choo, C.W., Detlor, B., Turnbull, D.: Information seeking on the Web: an integrated model of browsing and searching. *First Monday*, **5** (2000)
13. Guthrie, J.T., Mosenthal, P.: Literacy as multidimensional: locating information and reading comprehension. *Educ. Psychol.* **22**, 279–297 (1987)
14. Johnson, D., Meischke, H.: A comprehensive model of cancer-related information seeking applied to magazines. *Hum. Commun. Res.* **19**, 343–367 (1993)
15. Wilson, T.D.: The cognitive approach to information-seeking behaviour and information use. *Soc. Sci. Inf. Stud.* **4**, 197–204 (1984)
16. Marchionini, G.: *Information Seeking in Electronic Environments*. Cambridge University Press, New York (1995)
17. Dinet, J., Chevalier, A., Tricot, A.: Information search activity: an overview. *Revue Européenne de Psychologie Appliquée/Eur. Rev. Appl. Psychol.* **62**, 49–62 (2012)
18. Sharit, J., Hernández, M.A., Czaja, S.J., Pirolli, P.: Investigating the roles of knowledge and cognitive abilities in older adult information seeking on the Web. *ACM Trans. Comput.-Hum. Interact.* **15**, 1–25 (2008)
19. Sweller, J.: Cognitive load during problem solving: effects on learning. *Cogn. Sci.* **12**, 257–285 (1988)
20. Mayer, R.E., Heiser, J., Lonn, S.: Cognitive constraints on multimedia learning: when presenting more material results in less understanding. *J. Educ. Psychol.* **93**, 187–198 (2001)
21. Brunken, R., Plass, J.L., Leutner, D.: Direct measurement of cognitive load in multimedia learning. *Educ. Psychol.* **38**, 53–61 (2003)
22. Debue, N., Van De Leemput, C.: What does germane load mean? An empirical contribution to the cognitive load theory. *Front. Psychol.* **5**, 1099 (2014)
23. Galy, E., Cariou, M., Mélan, C.: What is the relationship between mental workload factors and cognitive load types? *Int. J. Psychophysiol.* **83**, 269–275 (2012)
24. Thatcher, A.: Web search strategies: the influence of Web experience and task type. *Inf. Process. Manag.* **44**, 1308–1329 (2008)
25. Compeau, D.R., Higgins, C.A.: Application of social cognitive theory to training for computer skills. *Inf. Syst. Res.* **6**, 118–143 (1995)
26. Bronstein, J.: The role of perceived self-efficacy in the information seeking behavior of library and information science students. *J. Acad. Librariansh.* **40**, 101–106 (2014)
27. Alexopoulou, P., Hepworth, M., Morris, A.: An investigation of multitasking information behavior and the influence of working memory and flow. In: *AIP Conference Proceedings*, pp. 37–43. AIP Publishing (2015)
28. Gwizdka, J., Chignell, M.: Individual differences and task-based user interface evaluation: a case study of pending tasks in email. *Interact. Comput.* **16**, 769–797 (2004)
29. Oswald, F.L., McAbee, S.T., Redick, T.S., Hambrick, D.Z.: The development of a short domain-general measure of working memory capacity. *Behav. Res. Methods* **47**, 1343–1355 (2015)
30. Gwizdka, J.: Assessing cognitive load on web search tasks. [arXiv:1001.1685](https://arxiv.org/abs/1001.1685) (2010)
31. Cheon, J., Grant, M.: Examining the relationships of different cognitive load types related to user interface in web-based instruction. *J. Interact. Learn. Res.* **23**, 29–55 (2012)

32. Amadiou, F., van Gog, T., Paas, F., Tricot, A., Mariné, C.: Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. *Learn. Instr.* **19**, 376–386 (2009)
33. Reeves, T.C., Benson, L., Elliott, D., Grant, M., Holschuh, D., Kim, B., Kim, H., Lauber, E., Loh, S.: Usability and instructional design heuristics for e-learning evaluation. In: *Proceedings of the 14th World Conference on Educational Multimedia, Hypermedia & Telecommunications*, pp. 1–9. Association for the Advancement of Computing in Education, Denver (2002)
34. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., van de Weijer, J.: *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, Oxford (2011)
35. Rayner, K.: Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **124**, 372–422 (1998)
36. Backs, R.W., Walrath, L.C.: Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Appl. Ergon.* **23**, 243–254 (1992)
37. Van Gerven, P.W.M., Paas, F., Van Merriënboer, J.J.G., Schmidt, H.G.: Memory load and the cognitive pupillary response in aging. *Psychophysiology* **41**, 167–174 (2004)
38. de Greef, T., Lafeber, H., van Oostendorp, H., Lindenberg, J.: Eye movement as indicators of mental workload to trigger adaptive automation. In: Schmorow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) *FAC 2009. LNCS (LNAD)*, vol. 5638, pp. 219–228. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02812-0_26
39. Marshall, S.P.: Identifying cognitive state from eye metrics. *Aviat. Space Environ. Med.* **78**, B165–B175 (2007)
40. Liu, H.-C., Lai, M.-L., Chuang, H.-H.: Using eye-tracking technology to investigate the redundant effect of multimedia web pages on viewers' cognitive processes. *Comput. Hum. Behav.* **27**, 2410–2417 (2011)
41. Di Stasi, L.L., Antolí, A., Gea, M., Cañas, J.J.: A neuroergonomic approach to evaluating mental workload in hypermedia interactions. *Int. J. Ind. Ergon.* **41**, 298–304 (2011)
42. Buettner, R.: Cognitive workload of humans using artificial intelligence systems: towards objective measurement applying eye-tracking technology. In: Timm, I.J., Thimm, M. (eds.) *KI 2013. LNCS (LNAD)*, vol. 8077, pp. 37–48. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40942-4_4
43. Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S.Z., Khawaji, A., Conway, D.: *Robust Multimodal Cognitive Load Measurement*. Springer, New York (2016). <https://doi.org/10.1007/978-3-319-31700-7>
44. Berka, C., Levendowski, D.J., Cvetinovic, M.M., Petrovic, M.M., Davis, G., Lumicao, M.N., Zivkovic, V.T., Popovic, M.V., Olmstead, R.: Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *Int. J. Hum.-Comput. Interact.* **17**, 151–170 (2004)
45. Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L.: EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Space Environ. Med.* **78**, 231–244 (2007)
46. Johnson, R., Popovic, D.P., Olmstead, R.E., Stikic, M., Levendowski, D.J., Berka, C.: Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biol. Psychol.* **87**, 241–250 (2011)
47. Bertolo, D., Dinet, J., Vivian, R.: Reducing cognitive workload during 3D geometry problem solving with an app on iPad. In: *2014 Science and Information Conference (SAI)*, pp. 896–900 (2014)

48. Sénécal, S., Fredette, M., Léger, P.-M., Courtemanche, F., Riedl, R.: Consumers' cognitive lock-in on websites: evidence from a neurophysiological study. *J. Internet Commer.* **14**, 277–293 (2015)
49. Borlund, P., Ingwersen, P.: The application of work tasks in connection with the evaluation of interactive information retrieval systems: empirical results. In: *Proceedings of MIRA 1999*, Glasgow, Scotland (1999)
50. Toms, E.G., O'Brien, H., Mackenzie, T., Jordan, C., Freund, L., Toze, S., Dawe, E., MacNutt, A.: Task effects on interactive search: the query factor. In: Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) *INEX 2007. LNCS*, vol. 4862, pp. 359–372. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85902-4_31
51. R Core Team: *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (2016)
52. West, B.T., Welch, K.B., Galecki, A.T.: *Linear Mixed Models: A Practical Guide Using Statistical Software*, 2nd edn. CRC Press, Boca Raton (2014)
53. Scharinger, C., Kammerer, Y., Gerjets, P.: Pupil dilation and EEG alpha frequency band power reveal load on executive functions for link-selection processes during text reading. *PLoS One* **10**, e0130608 (2015)
54. Crabb, M., Hanson, V.L.: Age, technology usage, and cognitive characteristics in relation to perceived disorientation and reported website ease of use. In: *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*. pp. 193–200. ACM, New York (2014)
55. Judd, T., Kennedy, G., Judd, T., Kennedy, G.: Measurement and evidence of computer-based task switching and multitasking by “net generation” students. *Comput. Educ.* **56**, 625–631 (2011)



A Pilot Study on Gaze-Based Control of a Virtual Camera Using 360°-Video Data

Jutta Hild¹(✉), Edmund Klaus¹, Jan-Hendrik Hammer¹, Manuel Martin¹, Michael Voit¹,
Elisabeth Peinsipp-Byma¹, and Jürgen Beyerer^{1,2}

¹ Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB),
Karlsruhe, Germany

jutta.hild@iosb.fraunhofer.de

² Vision and Fusion Laboratory, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Abstract. Over the last decades, gaze input appeared to provide an easy to use and less demanding human-computer interaction method for various applications. It appeared to be particularly beneficial in situations where manual input is either not possible or is challenging and exhausting like interaction with dynamic content in video analysis or computer gaming. In this contribution is investigated whether gaze input could be an appropriate input technique for camera control (panning and tilting) without any manual intervention. The main challenge of such an interaction method is to relieve the human operator from consciously interacting and to let them deploy their perceptive and cognitive resources completely to scene observation. As a first step, a pilot study was conducted operationalizing camera control by navigating in a virtual camera scene, comparing gaze control of the camera with manual mouse control. The experimental task required the 28 subjects (18 expert video analysts, 10 students and colleagues) to navigate in a 360° camera scene in order to keep track of certain target persons. Therefore, an experimental system was implemented providing virtual camera navigation in previously recorded 360fly camera imagery. The results showed that subjects rated gaze control significantly less loading than manual mouse control, using the NASA-TLX questionnaire. Moreover, the large majority preferred gaze control over manual mouse control.

Keywords: Virtual camera control · 360° video · Gaze input
Surveillance task

1 Introduction

Gaze-based control of user interfaces has been proposed and evaluated in plenty of contributions addressing various application domains. Researchers investigated gaze input for common desktop interaction tasks like object selection, eye typing, or password entry [1–3], for zooming maps or windows [4, 5], for foveated video streaming [6, 7], and PTZ camera remote control for surveillance [8] or teleoperation [9].

All implementations make use of gaze as a natural pointing »device«, as gaze is typically directed to the region of visual interest within the environment. Even though gaze has been evolved for perception, it has been shown that gaze can also be utilized

as a method for information input. Particularly, gaze input is an alternative in situations where manual input is not possible, e.g., due to motor impairment [2] and in hands-busy and attention switching situations [9]. Moreover, gaze input proved to be a beneficial alternative for interaction in dynamic scenes, e.g., for moving target selection in full motion video [10], where manual input might be exhausting and challenging.

Recently, the eye tracking manufacturer Tobii started to make eye tracking and gaze interaction suitable for another application domain where interaction in dynamic scenes happens – the mass market of computer gaming. They provide the low-cost eye tracking device Tobii »4C« for 149\$ (159€) [11]. Navigation in the scene of computer games using a first-person perspective is one of the proposed gaze input methods [12]. If the user directs their gaze, e.g., to the right corner of the current scene, the image section changes with the right corner subsequently becoming the next scene center. Thus, the visual focus of interest is brought to the scene center without any manual intervention. Such interaction models have also been proposed before by several authors investigating gaze input for computer gaming, e.g., [13, 14].

A similar kind of interaction is required when controlling a camera in a video surveillance task. Due to the rich visual input, this task can be very exhausting for the human operator, particularly, if the camera is mounted on a moving platform. Hence, any reduction of workload caused by less demanding human-computer interaction is welcome as it frees cognitive capacities for the actual surveillance task. A frequently occurring task is keeping track of a moving object, e.g., a person. If the object moves out of the currently displayed image section, the human operator must redirect the camera field of view. Gaze-based control of a camera appears compelling, keeping the camera focused on the object by just looking at this object. That way, the observer's visual attention could be focused on the (primary) surveillance task and the (secondary) interaction task is accomplished effortlessly at the same time.

In order to find out, whether such gaze interaction is appropriate and convenient, an experimental system was implemented simulating the control of a virtual camera as navigation in 360° video imagery. The system was evaluated in a user study with 28 participants, comparing gaze control versus manual control during the task of visually tracking a moving person.

2 Experimental System

The experimental system was implemented as a Java application which is able to play 360° video data recorded by the 360fly camera [15]. Figure 1 shows a video frame captured at an altitude of 30 m. For presentation to the user, the raw video data is rectified first, and in the next step, an image section of (width x height) $125^\circ \times 70^\circ$ of the rectified 360° video data is provided on the user interface (Fig. 2). In related work, Boehm et al. [16] introduce a similar system displaying an image section of a 185° fisheye camera.



Fig. 1. Video data captured by 360fly at an altitude of 30 m.

Gaze interaction is performed using the Tobii 4C eye-tracker [10]. They provide gaze data in different modes [17]; in our system, the »lightly filtered« mode is used and passing additional low-pass filtering before being processed in the application. Figure 3 shows the underlying gaze interaction model for navigation in the scene. When the gaze position is located within the center region (white), the displayed image section remains the same and the human operator is enabled to calmly inspect that central region. When the gaze is located off the center region (blue), the image section is re-centered on this gaze position. The farther the eye gaze is directed away from the center towards the edges or corners, the faster the image section is centered on the new gaze position; similar models have been proposed before for remote camera control for surveillance [8] and teleoperation [9]. Calculation of the repositioning speed is based on the squared Euclidian distance between current gaze position and screen center. The maximum allowed speed for image section repositioning (achieved if looking at the edges) is 3° per frame (frame rate is 60 Hz).



Fig. 2. Experimental system: image section displayed full-screen on a 14in laptop, equipped with a Tobii 4C eye-tracking device for gaze input, and a standard computer mouse providing the manual input alternative.

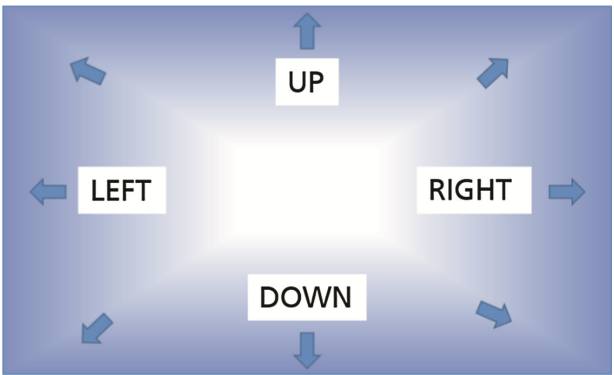


Fig. 3. The Gaze interaction model visualizing the activation dynamics on the screen: Gaze positions on the center region (white) have no effect. Gaze positions off the center region (blue) re-center the displayed image section to that gaze position; the closer to edges/corners (darker blue) a gaze position is located, the faster the image section is re-centered. (Color figure online)

The experimental system allows image section repositioning also with manual interaction using a computer mouse (Fig. 2). The user selects the image position of visual interest by pressing the left mouse button, then »drags« the image position with the left mouse button pressed to the wanted new position, for example the screen center.

3 Methodology

A pilot study was conducted to get first insight about the subjective workload of the gaze-based (virtual) camera control. 28 subjects (25 male, 3 female; 18 expert video analysts, 10 students and colleagues) performed the experimental task »Keep track of a person« using two different 3-min video sequences. Once, the test task instruction was to »Keep track of the person wearing the black jacket«, once »Keep track of the person wearing the red jacket« (Fig. 4). The video material was captured at an altitude of 30 m using a 360fly camera mounted on a 3DR solo drone [18]. The subjects were sitting at a distance of about 60 cm from the monitor (Fig. 5), the target persons' sizes therefore covered about $0.3^\circ \times 0.3^\circ$ of visual angle on screen.



Fig. 4. Screenshot of a test task with a target person. (Color figure online)

To ensure that subjects would have to reposition the scene in order to be able to keep track of the target person, the actors had been told to vary their motion trajectory and speed during video recording; thus, they temporarily moved straight on, or unpredictably, and sometimes shortly disappeared when walking under a tree. Furthermore, the drone and therefore also the camera trajectory carried out various motion patterns, like following an actor's trajectory, crossing an actor's trajectories, orbiting around the actors, or rotating at a stationary position. After performing the two test tasks, the subject answered the NASA-TLX [19, 20] questionnaire applied in the »Raw TLX« version, eliminating the weighting process.

For better interpretability of the NASA-TLX results for gaze input, the experimental design also incorporated performing the two test tasks using mouse input, and assessing it using the NASA-TLX. Half of the subjects performed the test tasks with gaze input



Fig. 5. Experimental setup.

first, the other half performed with mouse input first. The data recording of the 10 non-expert subjects was carried out in our lab, the data recording of the 18 expert video analysts was carried out at two locations of the German armed forces.

The procedure was as follows. Subjects were introduced into the experimental task but kept naïve in terms of the purpose of the investigation. Then, they performed the test tasks with the two interaction conditions one after another. In case of gaze input, subjects started performing the eye-tracker calibration provided by the Tobii-Software which requires fixating 7 calibration points; the calibration procedure was repeated until the offset between each fixated point and corresponding estimated gaze position was less than 1° of visual angle. Then, subjects got a different 3-min video sequence for training of the experimental task using that interaction technique. After that, subjects performed the two test tasks, immediately followed by rating their subjective workload using the NASA-TLX questionnaire. The mouse input condition was carried out performing the same three steps of training task, test tasks, and NASA-TLX rating. Finally, subjects were asked for their preferred interaction technique. The total duration of a session was about 30 min.

4 Results

The NASA-TLX results show that gaze input was rated with less workload both overall and in all single TLX categories. Results are provided using descriptive statistics as means with 1 standard deviation in Fig. 6 for all 28 subjects, in Fig. 7 for the expert video analysts only ($N = 18$). From those 18 experts, ten experts had much current

practice in video surveillance and therefore were analyzed again, separately; results are shown in Fig. 8.

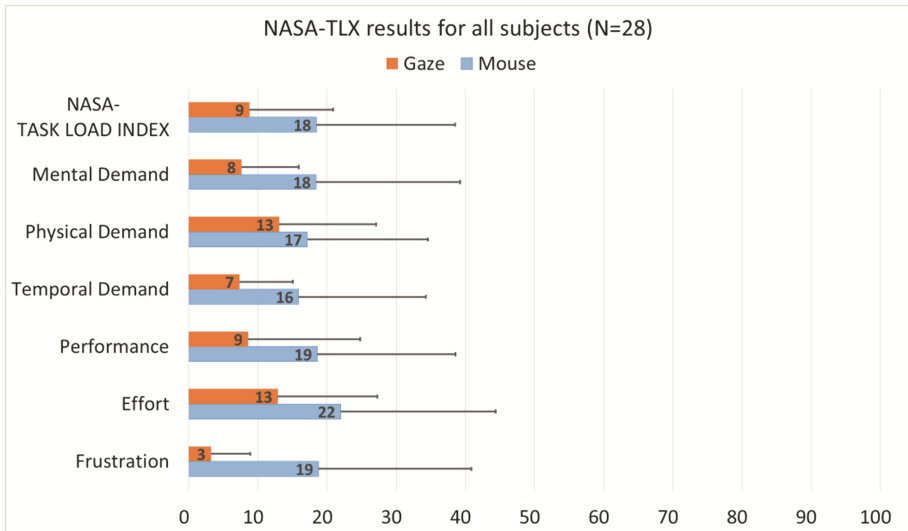


Fig. 6. Subjective workload with gaze input and mouse input, for all subjects.

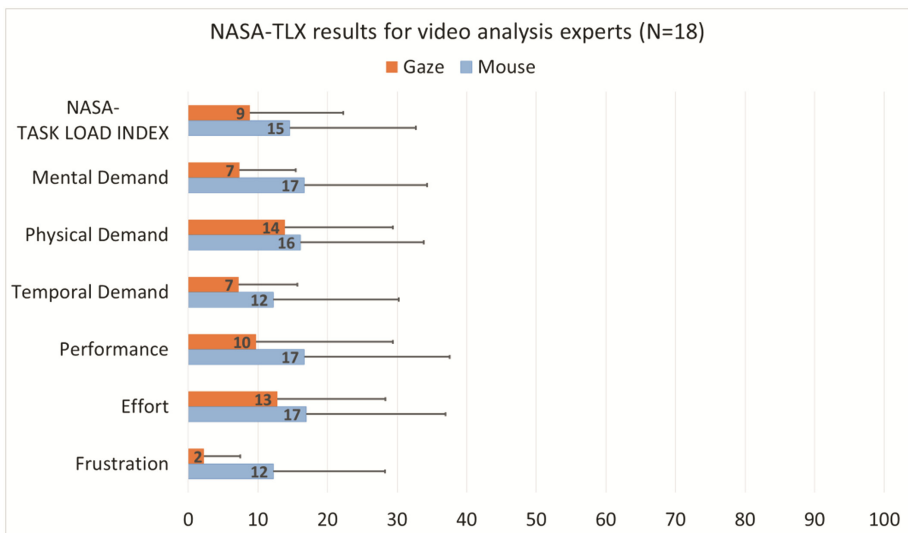


Fig. 7. Subjective workload with gaze input and mouse input, for subjects with expertise in video analysis.

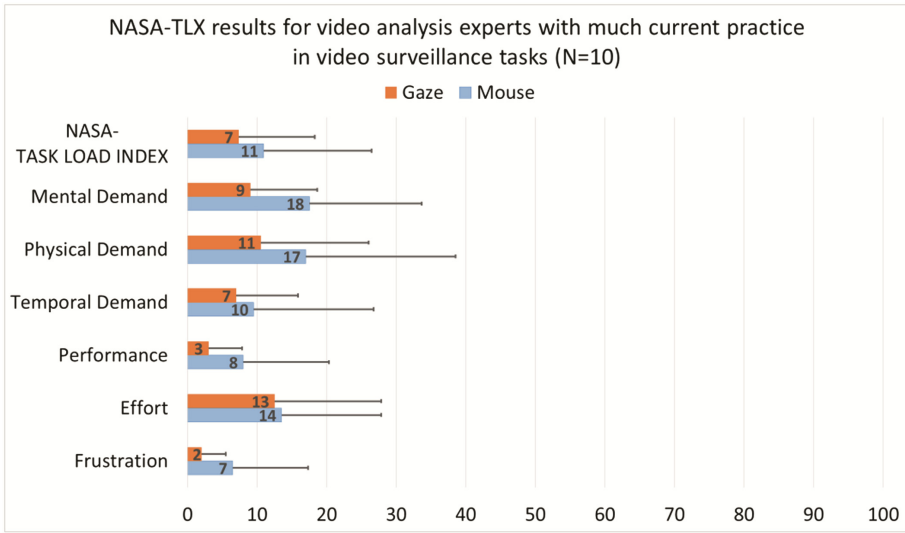


Fig. 8. Subjective workload with gaze input and mouse input, for subjects with expertise in video analysis and much current practice in video surveillance.

The NASA-TLX score is low for both interaction techniques, but it is significantly better for gaze input: A Wilcoxon signed-rank test for paired samples ($\alpha = 0.05$) revealed significant differences with $p < 0.001$ for $N = 28$, and $p < 0.05$ for $N = 18$; the result for the experts with much current practice in video surveillance ($N = 10$) is not significant ($p = 0.153$).

Analysis of the six subscales revealed further significant differences when analyzing all subjects ($N = 28$), for mental demand with $p < 0.05$, temporal demand $p < 0.01$, performance $p < 0.01$, effort $p < 0.05$, and frustration $p < 0.001$. Subscale analysis for expert video analysts ($N = 18$) and experts with much current practice in video surveillance ($N = 10$) still shows a significant difference between gaze and mouse for frustration ($p < 0.05$) despite the few data samples.

For mouse control, it can be observed that the subjective workload depends on video analysis expertise and current practice: The more expertise and practice, the lower the subjective workload (resulting in the NASA-TLX score difference between gaze and mouse being not significant any more, as reported above). However, for gaze control, subjective workload is very low for all subjects, independent of expertise. So, at least for control of a virtual camera, gaze input seems to be the more appropriate and convenient method to use.

Asked for their preference, 25 subjects preferred gaze input, 3 preferred mouse input ($N = 18$ experts: 15 preferred gaze input, 3 mouse input; $N = 10$ experts with much current practice in video surveillance: 10 preferred gaze input, 1 mouse input).

5 Conclusion

A pilot study was conducted in order to find out whether gaze input could be an appropriate input technique for camera control (panning and tilting) without any manual intervention. 28 subjects (18 expert video analysts from the German forces and 10 non-experts in video analysis) participated in the user study. Each performed the experimental task of tracking a target person using both gaze input and mouse input for navigation in a virtual camera, implemented based on 360° video imagery. The NASA-TLX showed that subjects rated both interaction conditions imposing rather little workload; however, gaze input was rated imposing significantly lower workload than mouse input. Hence, gaze input showed its potential to provide effortless interaction for this application, as it did for many other applications before.

Recently, the experimental system has been refactored and now besides navigation in recorded 360fly video data also allows live navigation in 360fly imagery. Future work will address gaze control for a real sensor, and user testing will show how workload would turn out to be in such condition with interaction latencies due to the necessary gimballed movements. Furthermore, future user studies will include more complex test tasks like observing more than one target object, as well as test tasks with a longer duration.

References

1. Jacob, R.J.: The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Trans. Inf. Syst. (TOIS)* **9**(2), 152–169 (1991)
2. Majaranta, P., Riih , K.J.: Twenty years of eye typing: systems and design issues. In: *Proceedings of the 2002 Symposium on Eye Tracking Research and Applications*, pp. 15–22. ACM, New York (2002)
3. Kumar, M., Garfinkel, T., Boneh, D., Winograd, T.: Reducing shoulder-surfing by using gaze-based password entry. In: *Proceedings of the 3rd Symposium on Usable Privacy and Security*, pp. 13–19. ACM, New York (2007)
4. Fono, D., Vertegaal, R.: EyeWindows: evaluation of eye-controlled zooming windows for focus selection. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 151–160. ACM, New York (2005)
5. Stellmach, S., Dachselt, R.: Investigating gaze-supported multimodal pan and zoom. In: *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 357–360. ACM, New York (2012)
6. Ryoo, J., Yun, K., Samaras, D., Das, S.R., Zelinsky, G.: Design and evaluation of a foveated video streaming service for commodity client devices. In: *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, New York (2016)
7. Illahi, G., Siekkinen, M., Masala, E.: Foveated video streaming for cloud gaming. *arXiv preprint [arXiv:1706.04804](https://arxiv.org/abs/1706.04804)* (2017)
8. Kotus, J., Kunka, B., Czyzewski, A., Szczuko, P., Dalka, P., Rybacki, R.: Gaze-tracking and acoustic vector sensors technologies for PTZ camera steering and acoustic event detection. In: *Workshop on Database and Expert Systems Applications (DEXA)*, pp. 276–280. IEEE (2010)
9. Zhu, D., Gedeon, T., Taylor, K.: “Moving to the centre”: a gaze-driven remote camera control for teleoperation. *Interact. Comput.* **23**(1), 85–95 (2010)

10. Hild, J., Kühnle, C., Beyerer, J.: Gaze-based moving target acquisition in real-time full motion video. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications, pp. 241–244. ACM, New York (2016)
11. Tobii Homepage. <https://tobiigaming.com/eye-tracker-4c/>. Accessed 05 Feb 2018
12. Tobii Homepage. <https://tobiigaming.com/games/assassins-creed-origins/>. Accessed 05 Feb 2018
13. Castellina, E., Corno, F.: Multimodal gaze interaction in 3D virtual environments. *COGAIN* **8**, 33–37 (2008)
14. Nacke, L.E., Stellmach, S., Sasse, D., Lindley, C.A.: Gameplay experience in a gaze interaction game. arXiv preprint [arXiv:1004.0259](https://arxiv.org/abs/1004.0259) (2009)
15. 360fly Homepage. <https://www.360fly.com/>. Accessed 05 Feb 2018
16. Boehm, F., Schneemilch, S., Schulte, A.: The electronic camera gimbal. In: AIAA Infotech@Aerospace Conference (2013)
17. Tobii Homepage. <https://tobii.github.io/CoreSDK/articles/streams.html>. Accessed 05 Feb 2018
18. 3DR Homepage. <https://3dr.com/solo-drone/>. Accessed 05 Feb 2018
19. Hart, S.G.: NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, no. 9, pp. 904–908. Sage Publications, Los Angeles (2006)
20. NASA TLX Homepage. <https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf>. Accessed 05 Feb 2018



Efficiency and User Experience of Gaze Interaction in an Automotive Environment

Benedikt Lux¹, Daniel Schmidl¹, Maximilian Eibl¹, Bastian Hinterleitner²,
Patricia Böhm¹, and Daniel Isemann¹(✉)

¹ Universität Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany
daniel.isemann@ur.de

² Audi Electronics Venture GmbH, Sachsstraße 20, 85080 Gaimersheim, Germany

Abstract. Over the last years the number and quality of information and entertainment systems in automobiles has been rising constantly. This presents the challenge to provide safe and user-friendly interaction techniques, the implementation of which can lead to a higher level of efficiency, safety and user experience. One novel and promising approach is to use the drivers gaze as input for interaction with infotainment systems. We implemented a virtual car interior model to test the efficiency and user experience of gaze interaction with automotive infotainment systems. In a user study with 20 participants we compared a gaze-based interaction style to a haptic interaction technique. The usability of both techniques turned out to be very similar, while the user experience and the efficiency varied in parts. We used an eye-tracking device to investigate gaze behavior, but due to some technical problems with the device our quantitative findings are not as reliable and robust as we would have hoped for and have to be interpreted with care. Our qualitative data indicated a preference for gaze interaction.

Keywords: Gaze interaction · Eye-tracking · Haptic interaction
In-vehicle information systems · User experience · Efficiency
Automotive environment

1 Introduction

Due to technical developments and new user demands, in-vehicle information systems (IVIS) have become more versatile and advanced. IVIS update traffic information, control the driver's comfort in the car and display multimedia content. This development in the car interior goes hand in hand with the growth of complexity of these systems [19]. Research has shown that this rise of complexity leads to a bigger mental workload for the driver especially when using poorly designed IVIS [14]. Inappropriately designed IVIS can not only cause minor driving errors like losing track of the lanes [26] or slower reaction times while driving [5] but are also among the main reason for traffic accidents in the United States [17]. With 8.8% for all fatalities in car accidents in the US in 2015 distraction-affected fatalities were the main reason for casualties in traffic.

One way to prevent distraction related accidents is the deployment of additional security systems in the car. A different approach is to improve the interaction with the IVIS while driving and hence minimize driver distraction. Therefore, many different design strategies have been explored over the last years, including larger infotainment screens and additional displays. In order to minimize driver distraction while interacting with IVIS we developed a gaze interaction based prototype and compared this novel interaction style with haptic control in a user study.

The prototype comprises a virtual car interior model and four infotainment system components that users could interact with: a navigation system, a speedometer, a climate control system and a telephone interface. In a within-subject study users had to complete typical scenarios (e.g. dismiss a call) with both test conditions (haptic vs. gazed based control) and usability and workload were assessed.

2 Related Work

There are many studies on the use of eye tracking systems in cars, for example for interacting with a car's dashboard [18] and for the regulation which supports the driver's awareness by warning him according to his eye movements [12]. Much research focuses on the cognitive workload of the driver and his driving performance. Beyond that it is crucial to examine the driver's user experience of IVIS [4]. For these reasons our user study assesses performance metrics as well as subjective measures. We use the System Usability Scale (SUS) and the User Experience Questionnaire (UEQ) to deepen the understanding about the user's attitude and reaction towards our system.

The approach of interacting with systems based on gaze direction like a mouse pointer is known to be an error-prone solution because of the so-called "Midas Touch" problem [10]. Because of this eye-tracking interaction alone is often not regarded as a robust way to communicate with systems not least because of the lack of technical precision and unpractical use of the gaze for interacting with systems, such as IVIS.

Consequently, most gaze-based interaction systems are combined with other modalities like speech and haptic systems. While auditory displays have been found to have the lowest impact on workload and driving performance it compounds the task efficiency [23] and appears to be slower [11]. Although the combination of eye tracking with touch has shown to be the fastest way to complete car dashboard tasks it is not the most practical one with regard to driver's reaction times and the feeling of being in control of the steering wheel. One convenient way for a multimodal gaze direction-based IVIS is the combination with a steering wheel button system. Kern et al. [11] show the advantages of this combined approach: It is possible to interact with screens which are hard or unpractical to reach while having the hands on the steering wheel.

For the evaluation of these multimodal interaction components a fundamental restriction is the limit of the eye-tracking system which requires a minimum angle for accuracy [6]. For the scientific purpose eye-tacker SMI RED the angle is 0.4° which is only reached with good calibration. Furthermore, there is a need for additional camera systems for eye detection across a wide range of the IVIS. Dobbstein et al. have shown

that a full landscape eye-tracking system still presents serious issues because of technical problems with multiple eye-tracking systems [4].

3 Experiment

3.1 Prototype

To explore the potential of gaze interaction in an automotive environment we implemented a virtual car interior model and four infotainment system components that users could interact with: a navigation system, a speedometer, a climate control system and a telephone interface. The model with the different emulated screens was presented on a wide TV screen to our study participants. The prototype contained one display for each component with several UI elements. Furthermore a number of tasks typical for in-car interaction (e.g. decline an incoming call from a friend on the phone display) were implemented. To compare gaze interaction to haptic interaction the prototype allowed two interaction styles: In the “eye tracking” condition the users could switch the displays by looking at them. In the “haptic condition” four specific buttons on the steering wheel could be selected to activate each display.

3.2 Setup

The experiment took place in the Future Interaction Lab of the University of Regensburg. In our setup the test persons were sitting in front of a 48” TV set with a distance from the screen of around 70 cm. The TV screen had a Tobii eye tracking device mounted underneath it. Between the users and the TV screen a gaming steering wheel was placed (see Fig. 1). Only the wheel’s hardware buttons were used during the experiment because there was no need for the users to steer. The participants were asked to perform eight tasks overall, four in each condition. One condition was using the eye tracker to select the displays on the screen, the other one was to select them by pushing a button.

3.3 Experiment Design

As dependent variables we assessed the Task Completion Rates and Times as well as subjective measures. The independent variables were the gaze interaction and haptic interaction. For our experiment, we chose a within-subject design, meaning that each of the participants had to perform tasks in both conditions. Half of the participants started with condition one, the other half with condition two in order to balance out learning effects.

In the beginning the participants were welcomed and introduced to our team. After a short explanation of the experiment, the eye tracker was calibrated. Afterwards, the prototype environment was started and the participant was advised to select the emulated displays by gazing at them, to familiarize themselves with the system and also to make sure the eye-tracker was calibrated correctly. Then the test coordinator read the tasks and the participant started with the execution. The time on task was assessed for each



Fig. 1. Test setup in front of a 48" TV with a steering wheel and an eye tracker beneath the TV.

task. After all tasks in the first condition, test users filled out several questionnaires, namely the System Usability Scale [22], the User Experience Questionnaire [13] and the NASA Task Load Index (NASA-TLX) [8] and were interviewed to gather qualitative data. The procedure was repeated with the second condition. In the end users were given a demographic questionnaire.

We used a mixed-methods approach, collecting qualitative and quantitative data to get a deeper understanding of the new interaction technique. Besides the interview items from the questionnaires mentioned before and additional demographic questions, we also measured Task Completion Times (TCT) and Task Completion Rates (TCR). The quantitative data was used to identify significant differences regarding efficiency, effectiveness or satisfaction whereas the qualitative data gave insights on the subjective behavior and perceived problems by the user.

3.4 Tasks

The test tasks were designed as interrupted tasks in which the user had to switch between the different displays two to four times, as it may occur in a real life scenario. To achieve this, we implemented events triggered by the action of the user. For example, the user changed the temperature in the car which triggered an incoming call on another display, which had to be declined. By using these interrupted tasks, it was possible to make users switch between displays during the test.

As shown in Fig. 2 four displays were presented on the screen. As is common in modern cars, the cruise control was positioned above the steering wheel in the Head-Up-Display (HUD). The navigation system was placed behind the steering wheel. In the center console we positioned the infotainment menu and the climate control. If selected by the user pressing one of the shoulder keys of the wheel or using gaze direction, the



Fig. 2. Display setup in the prototype with speed control, navigation system, infotainment system and climate control. (Color figure online)

displays were highlighted in red for providing visual feedback. Once selected the user was able to navigate inside this display by pressing the up and down button on the wheel. By pressing the “X” button on the wheel the user was able to select a button on the display.

The tasks consisted of small actions like the user having to start the navigation to a friend’s home address, setting the air conditioning to a specific level, setting the speed of the car’s cruise control or declining an incoming call.

3.5 Participants

13 male and 7 female subjects participated in the test, resulting in 20 subjects overall. The average age was 28, most participants were students at the University of Regensburg. Only one person did not have a driver’s license, and 13 owned a car. 14 users indicated that their technical affinity is off a normal to high level.

4 Results

As mentioned above, a mixed methods approach was chosen in recording the results of our user study. After every one of the 20 subjects had completed the user test with each interaction concept, they had to fill out some questionnaires. For the collection of quantitative data the metrics of the SUS and UEQ were selected to measure the level of

usability and user experience and the NASA-TLX to give an indication of the cognitive workload while using one of the two interaction techniques. The times users needed to complete a single task were recorded (task completion times, TCT). In addition to the quantitative data we also collected some qualitative data. Especially with rather small samples or samples that do not fully represent the whole target group, as in our case, it has been recommended to explain quantitative results with qualitative observations in a mixed-methods approach [21]. This should lead to a data analysis with greater explanatory potential [3] and help us answer the research questions whether the gaze interaction technique is more efficient than the haptic one and whether it records higher levels of usability and user experience. To gather such qualitative data the participants were asked questions about their experiences during the user test in short interviews. Furthermore, we added questions to our demographic questionnaire that give insights into the driving experience and the use of technical interaction components in cars.

4.1 System Usability Scale

The System Usability Scale (SUS) aims to identify users subjective rating of the usability of a tested system [2]. After rating the 10 oppositional word-pairs of the SUS on a Likert scale, the calculation of the SUS value leads to a number between 0 and 100 indicating the usability of a system. By admission of its inventors, the SUS represents a “quick and dirty solution” [2] which reveals only a rough tendency of the usability. With the SUS score of the haptic interaction ($M = 67.000$, $SD = 14.022$) and the score of the gaze interaction being not significantly different ($M = 65.625$, $SD = 14.478$), it is not possible to draw a comparative conclusion. To decide if these values are indicators for good or bad usability we compared them to SUS values from a larger range of tested systems, namely 2324 systems tested in a study by Bangor et al. [1] and 324 systems tested in a study by Sauro and Lewis [15] (Table 1). Compared to the Bangor et al. study the SUS values of our systems are below the mean and in the third highest quartile. Compared to the Sauro and Lewis study our SUS values are above both mean and median and in the second highest quartile. Hence, the score of the haptic and gaze interaction techniques are between the mean SUS Scores of those two studies. This may indicate an average level of usability.

Table 1. Basic information on SUS scores from the Bangor et al. and Sauro & Lewis studies.

| | Bangor et al. | Sauro & Lewis |
|--------------------------|---------------|---------------|
| N | 2324 | 324 |
| Mean | 70.1 | 62.1 |
| 1 st Quartile | 55.0 | 45.0 |
| Median | 75.0 | 65.0 |
| 3 rd Quartile | 87.5 | 75.0 |

A popular graphical outline of how to interpret the SUS score was created by Rauer [20]. The placement of the SUS results of our study according to this graphical interpretation scheme shows a similar assessment of our SUS scores (Fig. 3). However,

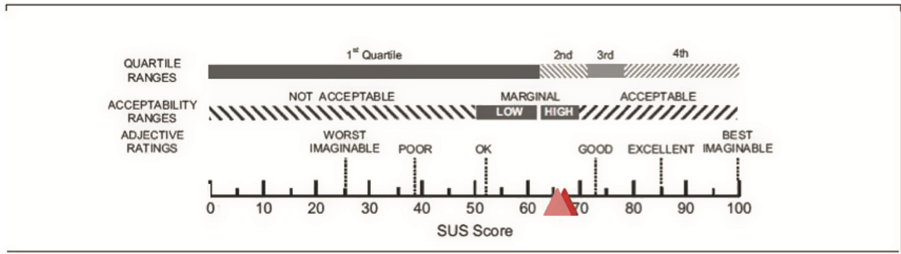


Fig. 3. Placement of the SUS scores of the gaze (light red) and haptic (dark red) interaction in Rauer's graphical interpretation scheme. (Color figure online)

Rauer's interpretation scheme is not referencing scientific findings or literature and should therefore be viewed with caution.

4.2 User Experience Questionnaire

As the SUS is only a metric to indicate a tendency of the level of usability we also used the User Experience Questionnaire (UEQ). The UEQ collects a wide range of data and allows for deeper and more detailed interpretation of the results. It consists of 26 items from six different dimensions and contains not only indicators for usability, but also for users' feelings and emotions while interacting with the system. We used the UEQ to measure the user experience of our two interaction concepts. Participants had to fill out the questionnaire and the analysis of the resulting data was done with a spread sheet, which is available for download at the UEQ online documentation [9].

The six main dimensions of the UEQ are *attractiveness*, *perspicuity*, *efficiency*, *dependability*, *stimulation* and *novelty*. While perspicuity, efficiency and dependability are indicators for ergonomic quality, stimulation and novelty show the hedonic quality. Attractiveness displays the overall impression of the research object [13].

Figures 4 and 5 show the UEQ mean scores per dimension of both interaction techniques. The UEQ score for attractiveness of the gaze interaction ($M = 1.02$, $SD = 0.68$) shows a higher value than the haptic interaction ($M = 0.52$, $SD = 0.91$). But this difference is not statistically significant ($t(20) = .056$, $p < .05$). The dimension of perspicuity shows a higher score for the haptic interaction ($M = 1.34$, $SD = 1.01$) than for the gaze interaction ($M = 0.76$, $SD = 0.90$) but is also not significant ($t(20) = .065$, $p < .05$). Efficiency shows almost equal UEQ scores for haptic interaction ($M = 0.93$, $SD = 0.68$) and for gaze interaction ($M = 0.91$, $SD = 0.63$) and is therefore not significant either ($t(20) = .953$, $p < .05$). A significant difference is found in the UEQ dimension dependability ($t(20) = .048$, $p < .05$). Here the haptic interaction technique shows a higher score ($M = 1.05$, $SD = 0.74$) than the gaze interaction technique ($M = 0.61$, $SD = 0.61$). On the other hand the gaze condition achieves a significantly ($t(20) < .001$, $p < .05$) higher score ($M = 1.26$, $SD = 0.62$) than the haptic condition ($M = 0.26$, $SD = 0.90$) in the dimension of stimulation. The largest significant difference ($t(20) < .001$, $p < .05$) and hence the possibly greatest advantage for the gaze interaction is seen in the dimension

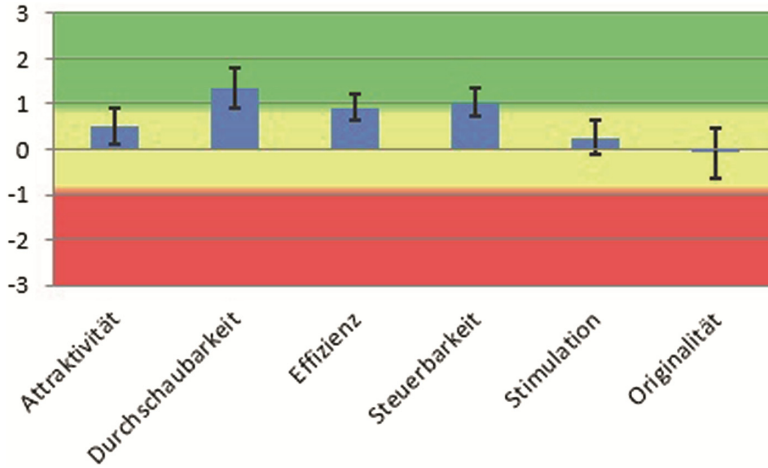


Fig. 4. UEQ scale of the six dimensions of the *haptic interaction* showing mean scores (blue bars) and confidence intervals (black lines). See text for English equivalents of the dimensions. (Color figure online)

of novelty. Here the score of the gaze condition ($M = 1,86, SD = 0.9$) is much higher than the score of the haptic condition ($M = -0.09, SD = 1.22$).

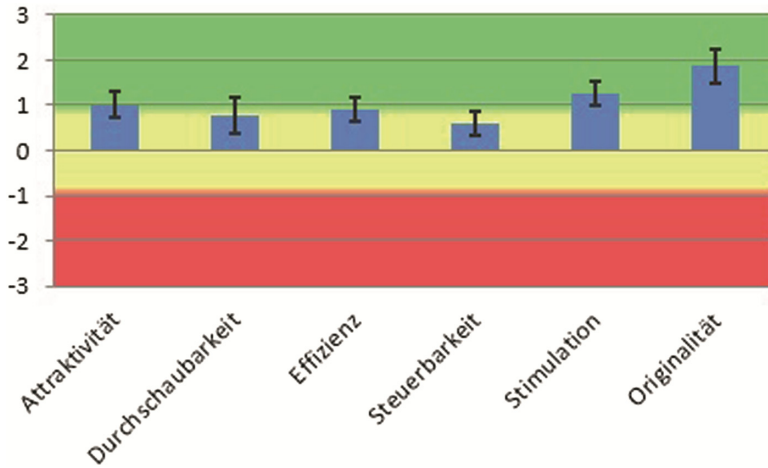


Fig. 5. UEQ scale of the six dimensions of the *gaze interaction* showing mean scores (blue bars) and confidence intervals (black lines). See text for English equivalents of the dimensions. (Color figure online)

To sum up the haptic interaction achieves better results in perspicuity (German: “Durchschaubarkeit”) and dependability (German: “Steuerbarkeit”), whereas the gaze interaction produces higher values for attractiveness (German: “Attraktivität”), stimulation (German: “Stimulation”) and novelty (German: “Originalität”). The dimension

of efficiency (German: “Effizienz”) is nearly equal for both interaction techniques. The only confidence interval which seems critical is the one of the novelty dimension of the haptic interaction. This value should not be over-interpreted because of possible misunderstandings with this item. Remarkable are especially the very high values for the originality of the gaze interaction and the perspicuity of the haptic one.

4.3 NASA – Task Load Index

To assess the workload experienced when using the two interaction techniques we used the well-established NASA-TLX. This task load index includes 6 items, measuring *mental*, *physical*, and *temporal demand*, the overall *effort*, *frustration level* and the subject’s satisfaction with their own *performance*. As Fig. 6 shows, the individual NASA-TLX dimensions for gaze and haptic interaction were comparable. To test for significant differences in the individual dimensions Wilcoxon tests for paired samples were used. The mean score of mental demand of the gaze interaction (8.35) was lower than the mean score of the haptic condition (9.1). The difference was not statistically significant ($z = -.692$; $p = 0.489$). The mean of the dimension physical demand for the gaze condition (6.7) was also lower compared to the haptic condition (7.5). The difference was not statistically significant ($z = -1.195$; $p = 0.232$). Regarding temporal demand the mean score for gaze (7.1) was higher than the one for haptic (6.4). The Wilcoxon test did not show a significant difference ($z = -7.51$; $p = 0.453$). The mean score for effort in gaze interaction (7.7) was very similar to the mean for haptic interaction (7.8). As expected the Wilcoxon test was not significant ($z = -2.65$; $p = 0.791$). Regarding performance the mean scores were relatively high with 14.1 for gaze and 13.1 for haptic interaction. The difference was not statistically significant ($z = -1.136$; $p = 0.256$). The mean frustration for gaze (6.75) was higher than the one for haptic

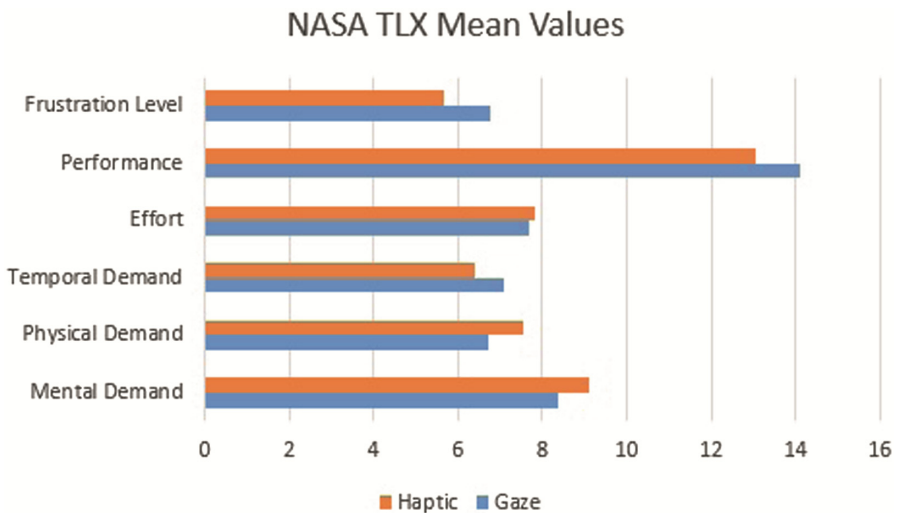


Fig. 6. The NASA-TLX item values show small differences.

interaction (5.6). The Wilcoxon test was not significant ($z = -6.79$; $p = 0.497$). As the mean scores of the dimensions were very similar and the Wilcoxon tests did not show any significant differences, the workload for gaze and haptic interaction would appear to be roughly equal.

4.4 Task Completion Rates

Every test person was able to complete every task using haptic and gaze input, so the task completion rate was 100% for both conditions.

4.5 Task Completion Times

The time a person needs to complete a task is a very important part of an interaction technique in an automotive environment. The more time a driver has to spend on completing a task with an interaction system, the greater the distraction from his primary task namely driving safely and focusing on the traffic. So measuring the TCT of every Task with both interaction techniques was indispensable.

Figure 7 illustrates the total task completion times for both conditions and shows that except for the first task, the haptic interaction was faster than the gaze interaction. To examine whether these differences are statistically significant we analyzed the average completion times and used Wilcoxon tests to detect significant differences. The mean task time for task 1 with gaze interaction (49.5 s) was lower than with haptic interaction (60.1 s), but the difference was statistically not significant ($z = -1,493$;

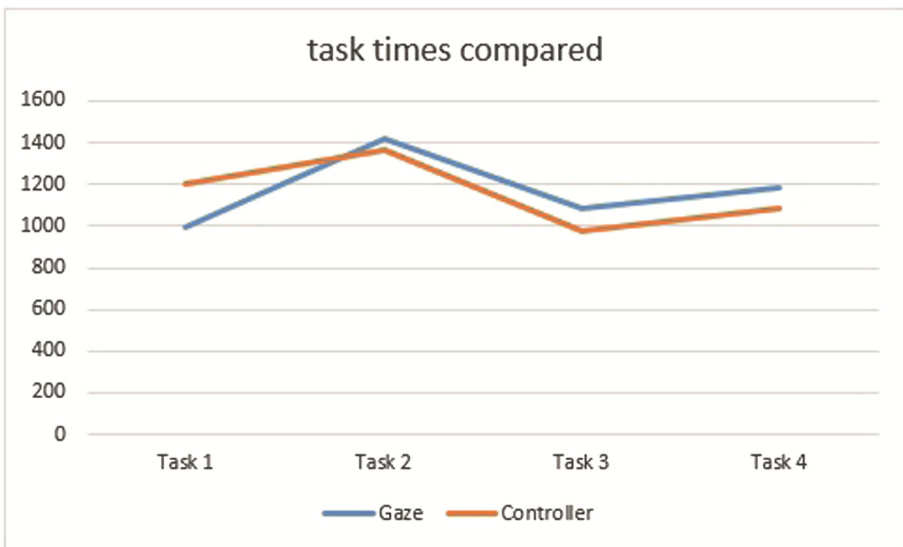


Fig. 7. The total of task completion times per task in seconds summed over all 20 participants. The comparison between haptic interaction (orange line) and gaze interaction (blue line) mostly shows shorter TCTs for the haptic interaction on average. (Color figure online)

$p = 0.135$). For task 2 the mean time on task for gaze (71 s) was higher than for haptic (68.2 s). The Wilcoxon test did not show a significant difference ($z = -5.6$; $p = 0.575$). The mean task time for task 3 was 53.3 s for gaze and lower for haptic with 48.7 s. The difference was statistically not significant ($z = -1.8$; $p = 0.07$). In Fig. 8 it can be seen, that the differences between conditions per subject are mostly small. Only for subjects 3, 4, 5, 11 and 17 there are more noticeable differences. It is perhaps noteworthy that the difference between the fastest and the slowest subject is nearly 350 s.

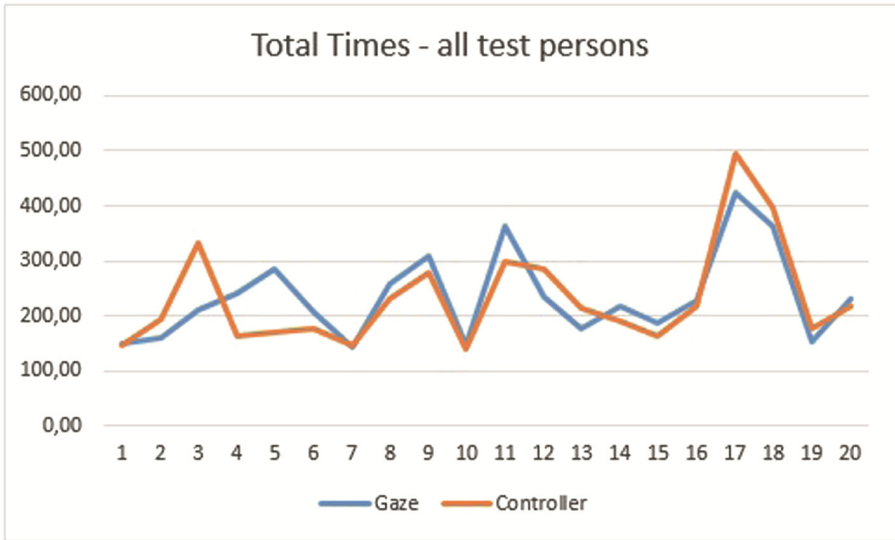


Fig. 8. The total completion time per person in seconds summed over all 4 tasks. The comparison between haptic interaction (orange line) and gaze interaction (blue line) shows mostly small differences between both interaction techniques. (Color figure online)

4.6 Structured Interview

After collecting the quantitative data from all subjects under both research conditions the subjects were asked to answer some short questions about their experiences and impressions during the experiment. The goal of this short structured interview was to get answers that are succinct and easy to compare. The potential drawbacks which such an interview method entails, such as incentivizing certain answers, were kept in mind and mitigated as far as possible [16].

The interview comprised seven free-text questions:

1. Which interaction technique did you find more pleasing and why?
2. Could you briefly compare both techniques?
3. Did you find the separation of the four displays intuitive?
4. Would you rather prefer interacting with one display?
5. Can you imagine using gaze interaction in a car?

6. Which concerns do you have about gaze interaction?
7. Do you have other remarks to both techniques?

These questions were appended with two Likert scale questions, which inquired how intuitive switching between displays was with each technique:

8. How intuitive did you find switching between the displays with gaze interaction?
9. How intuitive did you find switching between the displays with haptic interaction?

The answers to these questions were collected and afterwards structured by a qualitative content analysis. The analyzed answers were assigned to the inductively developed categories which were counted based on how often they occurred.

Question 1

14 of the 20 subjects answered that they would prefer the gaze interaction. 4 subjects would rather choose the haptic interaction. The remaining 2 gave no usable answer.

Question 2

The answers to the second question revealed four perceived advantages of the gaze interaction: The technique was seen as innovative (mentioned 4 times), faster (2), less stressful (2) and more intuitive (1).

Question 3

17 subjects found the separation and placement of the four displays intuitive. 2 had the opposite opinion and 1 did not give a clear answer.

Question 4

Only 3 subjects would prefer to interact with a single display, 16 would not and 1 answer was not usable.

Question 5

All suspects gave a clear answer if they could imagine gaze interaction in a car. 3 could not, 17 subjects said, they could.

Question 6

The purpose of the final free-text question was to generate further thoughts or ideas by the subjects. A noteworthy remark was made by participant 10, who opined that a car manufacturer could outdo competitors by using gaze-interaction in a car: "You could surely score bonuspoints against your competition, if you used gaze interaction (Translated from German: "Mit einer Blickinteraktion könnte man sicher gegenüber den Wettbewerbskonkurrenten punkten"). Subject 18 who was completely against both systems and preferred haptic hardware buttons for discrete functions "I find it cumbersome to first select a display in order to then interact with something there. I would rather control things directly" (Translated from German: "Ich finde es umständlich erst ein Bildschirm auszuwählen, um dann dort etwas zu bedienen. Ich würde lieber die Dinge direkt steuern").

5 Interpretation

The SUS scores of both interaction techniques are around the mean value of the compared 2648 scores. They do not differ much from one another, so there is no clear difference between the two systems. The UEQ data by contrast reveals some differences. Taking a closer look at Figs. 4 and 5, it would appear that the two patterns are nearly axis-symmetric. The gaze interaction shows good results in the dimensions that the haptic interaction does not and vice versa. If one could improve the scores of the ergonomic dimensions of the gaze interaction, one could improve the whole UEQ result and it would probably also lead to a higher SUS score. The replies from the subjects' interviews about the concerns of the experienced gaze-interaction indicated that there were several problems with the Tobii EyeX eye tracking device during the experiment. Several times the tracked gaze was lost or too imprecise. This caused longer TCTs and with the TCTs being negatively correlated to the ergonomic UEQ items, the gaze interaction showed a bad performance [24]. This might also be a reason for worse UEQ and SUS results. Repeating the experiment with more robust eye tracking functionality would surely lead to better results for the gaze interaction.

This interpretation also explains why most of the subjects could imagine to use a gaze interaction system in a car, although the quantitative test results do not reveal a big advantage for this interaction technique.

Also the TCTs of Task 1, the only task where gaze interaction was faster than haptic interaction, support this theory, because it was the task with the shortest distance for the users gaze movements and therefore had the smallest error rate for the eye tracker.

A closer look at the dataset of subject 11 also underlines this theory. His individual SUS scores for gaze (SUS score: 50) and haptic interaction (SUS score: 75) showed the greatest difference among all subjects. He was one of the few people who rated gaze interaction badly in the UEQ dimensions efficiency, stimulation and novelty. His TCTs were all in all about 100 s slower with gaze interaction than with haptic interaction. To question b) from the structured interview he answered, that the eye tracking was very inconsistent and imprecise, so that he interacted with the wrong display regularly. The interaction with the steering wheel might have been much easier, because using the gaze interaction had been too inconsistent and imprecise. He had regularly interacted with the wrong widget. In spite of all this he could imagine to use the gaze interaction in a car, because it would be more exciting if it worked better.

6 Conclusion

In our user study we wanted to find the answer to two research questions. The first was if a gaze interaction technique in an automotive environment is more efficient than a haptic one. To answer this question one could point to the average or total TCTs and come to the conclusion that the gaze interaction technique is not more efficient. However a closer look at the times of each task shows, that for example the first task was completed faster using gaze interaction. Also 9 out of 20 subjects were faster with the eye-tracker. Especially subject 17 and 18, the two elderly subjects (>70 years), who

had to completely learn both techniques from scratch, had more problems with the haptic interaction and were faster with the gaze interaction. If we keep the problems with the eye-tracker in mind, the overall results could possibly have been better. Even the subjective efficiency of the suspects was not significantly different, as the UEQ results indicated.

The second research question was, if the eye-tracking system is more user-friendly. Therefore we wanted to estimate the usability and user experience of both interaction styles. The SUS indicated almost no difference in usability with the two scores being very similar (SUS of gaze interaction: 65.625, SUS of haptic interaction: 67). The results of the UEQ demonstrated that both techniques offer a completely different experience of interaction. Almost every dimension of the UEQ is good in one interaction technique and bad in the other one. Where the gaze interaction scores in the hedonic dimensions of attractiveness, stimulation and novelty, the haptic interaction points to the ergonomic dimensions perspicuity and dependability. Only the dimension of the efficiency is nearly equal for both techniques. The interview showed clearly that 14 out of the 18 usable answers of subjects to the question which system they would rather use, expressed preference for gaze interaction.

Overall our study cannot provide a completely accurate answer to our research questions. While our quantitative data shows a small advantage for the haptic interaction, the qualitative data points in the opposite direction and the hardware problems mentioned above may even have hampered a clearer endorsement of gaze interaction.

The variation in distances between different subjects and the eye tracker as well as the precision of the eye tracker itself caused problems with the gaze estimation. In addition, some subjects had more problems than others because of their height or their glancing behavior. It was obvious that corresponding to the two glancing stereotypes of Fridman et al. [7], the eye tracking device had more problems with so-called *owls* and less problems with so-called *lizards*.

Despite these restrictions we could identify some positive tendencies and aspects of gaze interaction. Therefore it may be worth trying to pursue the development of this interaction technique further. A key factor would be the use of a more stable and precise eye-tracker. There is a high chance that this would improve the results of future studies researching gaze interaction. The statements made in the interviews clearly showed that gaze interaction holds some potential as an interaction technique in future cars.

References

1. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. *Int. J. Hum.-Comput. Interact.* **24**(6), 574–594 (2008)
2. Brooke, J.: SUS – a quick and dirty usability scale. *Usability Eval. Ind.* **189**(194), 4–7 (1996)
3. Creswell, J.W., Creswell, J.D.: *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. Sage Publications, Thousand Oaks (2017)
4. Dobbstein, D., Walch, M., Köll, A., Sahin, Ö., Hartmann, T., Rukzio, E.: Reducing in-vehicle interaction complexity: gaze-based mapping of a rotary knob to multiple interfaces. In: *Proceedings of the 15th International Conference of Mobile and Ubiquitous Multimedia*, pp. 311–313. ACM, December 2016

5. Donmez, B., Boyle, L.N., Lee, J.D.: Safety implications of providing real-time feedback to distracted drivers. *Accid. Anal. Prev.* **39**(3), 581–590 (2007)
6. Feit, A.M., Williams, S., Toledo, A., Paradiso, A., Kulkarni, H., Kane, S., Morris, M.R.: Toward everyday gaze input: accuracy and precision of eye tracking and implications for design. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1118–1130. ACM, May 2017
7. Fridman, L., Lee, J., Reimer, B., Victor, T.: ‘Owl’ and ‘Lizard’: patterns of head pose and eye pose in driver gaze classification. *IET Comput. Vis.* **10**(4), 308–314 (2016)
8. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: *Advances in Psychology*, North-Holland, vol. 52, pp. 139–183 (1988)
9. Hinderks, A.: UEQ Online: Downloads. <http://www.ueq-online.org/index.php/user-experience-questionnaire-download/?lang=de>. Accessed 17 Jan 2015
10. Jacob, R.J.: What you look at is what you get: eye movement-based interaction techniques. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 11–18. ACM, March 1990
11. Kern, D., Mahr, A., Castronovo, S., Schmidt, A., Müller, C.: Making use of drivers’ glances onto the screen for explicit gaze-based interaction. In: *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 110–116. ACM, November 2010
12. Langner, T., Seifert, D., Fischer, B., Goehring, D., Ganjineh, T., Rojas, R.: Traffic awareness driver assistance based on stereovision, eye-tracking, and head-up display. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3167–3173. IEEE, May 2016
13. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (ed.) *USAB 2008*. LNCS, vol. 5298, pp. 63–76. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89350-9_6
14. Lee, J.D., Caven, B., Haake, S., Brown, T.L.: Speech-based interaction with in-vehicle computers: the effect of speech-based e-mail on drivers’ attention to the roadway. *Hum. Factors* **43**(4), 631–640 (2001)
15. Lewis, J.R., Sauro, J.: The factor structure of the system usability scale. In: Kurosu, M. (ed.) *HCD 2009*. LNCS, vol. 5619, pp. 94–103. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02806-9_12
16. Mayring, P.: Qualitative inhaltsanalyse. In: Mey, G., Mruck, K. (eds.) *Handbuch Qualitative Forschung in der Psychologie*, pp. 601–613. VS Verlag für Sozialwissenschaften, Wiesbaden (2010). https://doi.org/10.1007/978-3-531-92052-8_42
17. National Highway Traffic Safety Administration: 2015 motor vehicle crashes: overview. *Traffic safety facts research note*, pp. 1–9 (2016)
18. Poitschke, T., Laquai, F., Stamboliev, S., Rigoll, G.: Gaze-based interaction on multiple displays in an automotive environment. In: *2011 IEEE International Conference Systems, Man, and Cybernetics (SMC)*, pp. 543–548. IEEE, October 2011
19. Purucker, C., Naujoks, F., Prill, A., Neukum, A.: Evaluating distraction of in-vehicle information systems while driving by predicting total eyes-off-road times with keystroke level modeling. *Appl. Ergon.* **58**, 543–554 (2017)
20. Rauer, M.: Quantitative Usability-Analysen mit der System Usability Scale (SUS). *Nachrichten, Tipps & Anleitungen für Agile, Entwicklung, Atlassian Software (JIRA, Confluence, Stash, ...)*. Seibert Media (2011). <https://blog.seibert-media.net/blog/2011/04/11/usability-analysen-system-usability-scale-sus/>. Accessed 30 Mar 2017

21. Ritchie, J., Lewis, J., Nicholls, C.M., Ormston, R. (eds.): *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. Sage, Thousand Oaks (2013)
22. U.S. Department of Health & Human Services: System Usability Scale (SUS) (2013). <http://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html>. Accessed 8 May 2015
23. Weinberg, G., Harsham, B., Medenica, Z.: Evaluating the usability of a head-up display for selection from choice lists in cars. In: *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 39–46. ACM, November 2011
24. Zwahlen, H.T., Adams, C.C., DeBals, D.P.: Safety aspects of CRT touch panel controls in automobiles. In: *Vision in Vehicles II. Second International Conference on Vision in Vehicles*, pp. 335–344 (1988)



Investigation of Factors Affecting the Usability Evaluation of an Adaptive Cruise Control System

Akihiro Maehigashi¹(✉), Kazuhisa Miwa¹, Hirofumi Aoki²,
and Tatsuya Suzuki³

¹ Graduate School of Informatics, Nagoya University, Nagoya, Japan
mhigashi@cog.human.nagoya-u.ac.jp

² Institute of Innovation for Future Society (MIRAI), Nagoya University,
Nagoya, Japan

³ Graduate School of Engineering, Nagoya University, Nagoya, Japan

Abstract. In this study, we investigate the factors affecting the usability evaluation of an adaptive cruise control (ACC) system. In this experiment, the participants drove a Toyota Prius car with an ACC on a highway. We sampled 215 types of driving data recorded at a frequency of 60 Hz during driving. At each of the six designated stop points on the driving course, the participants stopped their cars and evaluated the usability of the ACC system by answering the usability questionnaire for automation systems. The participants' driving styles were measured using the driving style questionnaire. The multiple regression analyses showed that the participants' driving styles, the ACC's driving control, and the participants' intervention in the driving control of the ACC influenced the usability evaluation. The results were discussed in terms of the human-automation interactions and the design principles of an ACC.

Keywords: Adaptive cruise control · Usability · Driving style
Driving control · Automation system

1 Introduction

1.1 ACC and Usability

Automation systems, such as cleaning robots and automated driving systems, are becoming very popular in recent years because of the technological advances in this century. An automation system is a technology that autonomously behaves on behalf of humans [1]. In particular, the development and the prevalence of highly advanced automated systems have been remarkable. The Society of Automotive Engineers International [2] defines six levels of automation from complete manual driving to fully automated driving. In this year (i.e., 2018), automation systems at level 1 are being commercially produced. A level 1 system in a vehicle can sometimes assist the human driver with either steering the vehicle or with braking and accelerating it but not both simultaneously. In level 2 systems, the vehicle can itself actually control both the

steering and the braking and accelerating simultaneously under some circumstances; however, these systems are still in their nascent stage.

Adaptive cruise control (ACC) is a system that can assist the human driver with braking and accelerating, and this is a level 1 system. The ACC senses the vehicle in front by using a radar sensor and controls the vehicle speed by keeping a certain distance between the two vehicles; the distance is determined by the driver. Also, when the vehicle in front is not sensed, the ACC maintains the speed set by the driver. Although the ACC has already been developed, usability evaluation is essential for these automated systems [3].

Usability tests are performed to develop and improve home electric appliances and information technology devices. Usability is defined as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” [4]. Effectiveness is “the accuracy and completeness with which users achieve specified goals.” Efficiency is “the resources expended in relation to the accuracy and completeness with which users achieve goals.” Satisfaction is “the freedom from discomfort and positive attitudes toward the use of the product” [4].

A usability test is an evaluation method of artifacts and interfaces [5]. In the objective evaluation of usability, the measurable indexes should be determined according to the definitions of effectiveness, efficiency, and satisfaction; subsequently, human behaviors are measured based on the indexes [6]. For example, effectiveness, efficiency, and satisfaction could respectively be measured by the percentage of errors that occur during a task, the task completion time, and the frequency of using an artifact. Moreover, the subjective evaluation of usability is performed throughout the user experience, and this gives the designers helpful information for improving the user experience [6]. The subjective evaluation is usually done by responding to questionnaires and interviews.

Automation systems have features that cognitive artifacts, such as computers, do not have [3]. Therefore, it is difficult to evaluate the usability of the systems based on the traditional elements of usability, i.e., effectiveness, efficiency, and satisfaction. Maehigashi et al. [3] developed a usability evaluation questionnaire for automation systems. In this questionnaire, the following three new elements were added to evaluate the automation system usability: understandability, discomfort, and motivation. Understandability is the comprehensibility of the intentions of automation systems. Discomfort is the unpleasant feeling associated with behaviors of the automation systems; these are independent of task performance. Motivation is the desire of users to perform tasks by themselves.

1.2 Purpose

The purpose of this study is to investigate the factors that determine the usability evaluation of ACC. This investigation is important for understanding the features of a human–automation system interaction and for improving the usability of an ACC.

The factors to be investigated were driving control and driving style factors (see Fig. 1). The driving factor was divided into the following three categories: ACC driving control, ACC-driver driving control, and driver driving control. The ACC

driving control factor indicated the ACC's driving control or the results of the control when the ACC was activated. The ACC-driver driving control factor referred to the driver's driving control or the results of the control when the ACC was activated. Also, the driver driving control factor expressed the driver's driving control or the results of the control when the ACC was deactivated.

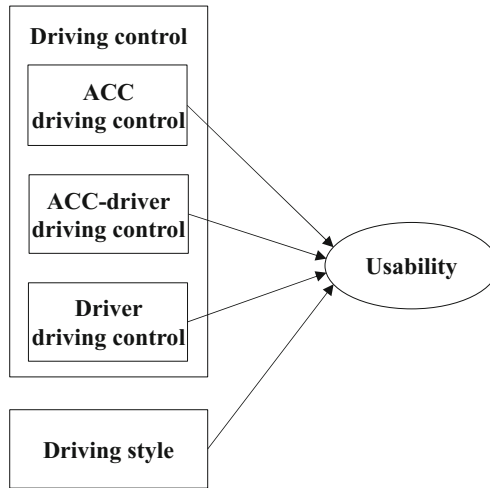


Fig. 1. Three types of driving control factors and driving style factor.

Moreover, previous studies about the usability evaluation showed that the personal traits of users influence the usability evaluation of various kinds of systems [7]. Therefore, we decided to investigate the driving style factor because the driver's driving style was considered to influence the usability evaluation of an ACC.

Most of the previous studies for ACCs used driving simulators e.g., [8]. A few studies used real cars with ACCs e.g., [9]. Furthermore, very few studies collected detailed driving data and discussed the ACC. In this study, we conducted an experiment with an actual car that was equipped with an ACC and investigated the usability evaluation of an ACC based on detailed driving data.

2 Experiment

In the experiment, the participants drove a Toyota Prius car with an ACC, which was called a radar cruise control system. The driving data was collected using the Vector CANcardXL interface. We collected 215 types of driving data at 60 Hz while driving; this included data, such as the activation and the deactivation of the ACC, the vehicle speed, the steering angle, the accelerator opening degree, and the brake hydraulic pressure. Moreover, to measure a driver's driving style, we used a driving style questionnaire that comprised 18 questions about eight elements [10]. To assess the

usability of the ACC, we used a usability questionnaire for the automation system that comprised 18 questions about six elements [3].

2.1 Participants and Procedure

Twenty persons (13 males and 7 females) participated in this experiment. The mean age was 41.95 years (ranging from 21 to 55 years). The mean experience as a driver was 21.95 years (ranging from 2 to 34 years). All the participants drove on a daily basis; however, they did not have any experience in driving with an ACC.

First, the participants answered the driving style questionnaire and were informed about the driving course. The driving course comprised six driving sections (see Fig. 2). The total length of the course was about 80 km. Next, the functions and the operation method of the ACC were explained to the participants. They were instructed to use the ACC as much as possible during driving. After that, they started to drive.

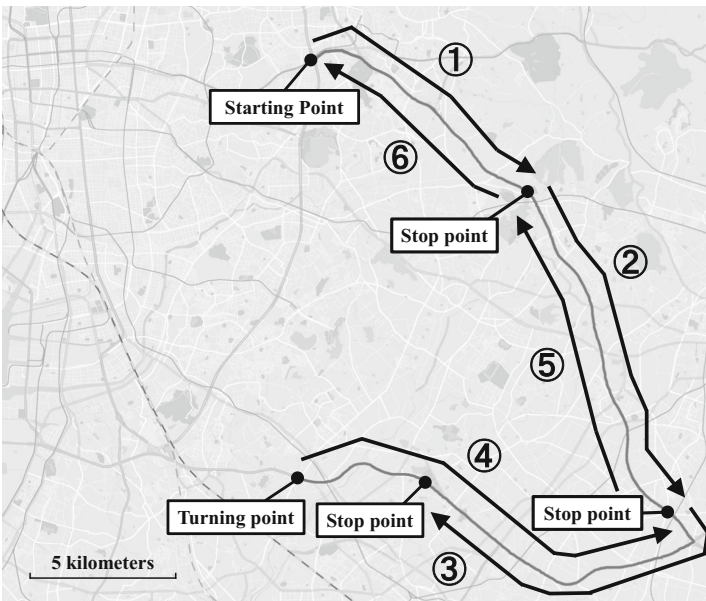


Fig. 2. Driving course

While driving, the experimenter was seated in the front passenger seat. When there was no vehicle ahead, the ACC maintained a speed of 100 km/h, which was set up by participants at the beginning of each section. When there was a vehicle ahead, the participants freely chose a preferred distance from three options as the distance between their car and the vehicle ahead. Furthermore, the participants were allowed to ask questions about the operations of the ACC anytime during the driving.

At the end of each section, the participants stopped the car and evaluated the usability of the ACC by answering the usability questionnaire. After a 5 min break,

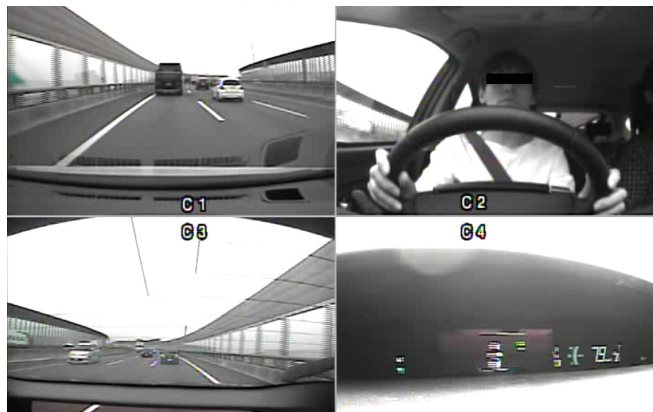


Fig. 3. Still from video recordings during driving. Camera 1 monitored the road ahead. Camera 2 monitored the driver. Camera 3 monitored the road behind. Camera 4 monitored the ACC dashboard indicator, which displayed the status of the ACC.

they restarted their drive. After driving all the six sections, they were interviewed about their interaction with the ACC. Figure 3 shows a still from the video recordings during the driving experiment.

2.2 Result

Analysis Method. In each section, the average driving time was 518.36 s, while the average time during ACC driving control was 404.19 s. Also, the average time during ACC-driver driving control was 21.19 s. Finally, the average time of the driver driving control was 92.97 s.

To investigate the determinative factors, we conducted a multiple regression analysis. The ACC is a system that controls acceleration and braking; therefore, we chose 20 independent variables related to accelerating and braking from the driving data for analysis (Table 1). To detect multicollinearity, we calculated variance inflation factor (VIF) from the correlation coefficients. As a result, the VIF between the number of accelerations and the number of decelerations in the ACC driving control factor was more than 10 ($VIF = 65.21$). Therefore, the number of decelerations was eliminated from the data; consequently, 19 independent variables were used for the analyses. Also, eight out of the 19 independent variables, which were the average scores of each element of the driving style questionnaire, were extracted from the answers. We calculated the average scores of the answers to the two questions for each element in the driving style questionnaire.

The dependent variables were the average scores of the answers to the three questions for each element in the usability questionnaire. Finally, we conducted multiple regression analyses of all the independent variables for each of the dependent variables (Table 2).

Table 1. Independent variables for three types of driving control factors and driving style factor and the corresponding explanations of the variables.

| Factor | Independent variables | Explanation of variables |
|----------------------------|---|--|
| ACC driving control | Driving time | The duration for which the ACC controlled the driving speed |
| | Number of times of acceleration | The number of times when the ACC continued to accelerate more than 1 s |
| | Number of times of deceleration | The number of times when the ACC continued to decelerate more than 1 s |
| | Rapid acceleration | The maximum value of the acceleration |
| | Rapid braking acceleration | The maximum value of the braking acceleration |
| | Intervehicular distance variation | The variance of the intervehicular distance |
| ACC-driver driving control | Driving time | Duration for which the drivers stepped on the accelerator |
| | Number of times of stepping on accelerator | The numbers of times when the drivers stepped on the accelerator |
| | Number of times of stepping on the brake | The numbers of times when the driver stepped on the brake |
| | Rapid acceleration | The maximum value of the rapid acceleration |
| | Rapid braking acceleration | The maximum value of the rapid braking acceleration |
| | Intervehicular distance variation | The variance of the intervehicular distance |
| | Number of times of switching (ACC→Driver) | The numbers of times when the driver switched off the ACC |
| Driver driving control | Driving time | Duration when the driver stepped on the accelerator |
| | Numbers of times of stepping on accelerator | The numbers of times when the driver stepped on the accelerator |
| | Numbers of times of stepping on the brake | The numbers of times when the driver stepped on the brake |
| | Rapid acceleration | The maximum value of the rapid acceleration |
| | Rapid braking acceleration | The maximum value of the rapid braking acceleration |
| | Intervehicular distance variation | The variance of the intervehicular distance |
| | Number of times of switching (Driver→ACC) | The numbers of times when the driver switched on the ACC |
| Driver's driving style | Confidence in driving | Existence of confidence in driving skills |
| | Passive mentality when driving | Reluctance in driving |
| | Impatient driving tendency | Tendency to drive in a rushed manner |
| | Scrupulous driving tendency | Tendency to drive methodically |
| | Signal-oriented preliminary preparing driving | Tendency to prepare for traffic signal |
| | Driving as a status symbol | Considering driving as status symbol |
| | Unstable driving tendency | Tendency to drive unstably |
| | Natural worrier | Tendency to drive cautiously |

Table 2. Results of the multiple regression analyses. I, II, III, IV, V, and VI represent effectiveness, efficiency, satisfaction, understandability, discomfort, and motivation respectively. The values indicate the standard regression coefficients. ⁺ $p < .10$, ^{*} $p < .05$, ^{**} $p < .01$, ^{***} $p < .001$

| Factor | Independent variable | I | II | III | IV | V | VI |
|----------------------------|--|---|--------|--------|--------|--------|---------|
| ACC driving control | Driving time | | | | | | |
| | Number of times of acceleration | | | ***.33 | ** .30 | | |
| | Rapid acceleration | | | | | | ***-.30 |
| | Rapid braking acceleration | | | | *.23 | | |
| | Intervehicular distance variation | | ** .27 | | | *-.18 | ***-.32 |
| ACC-driver driving control | Driving time | | **-.33 | | | | |
| | Number of times of stepping on accelerator | | ***.40 | | | | |
| | Number of times of stepping on the brake | | | | | | |
| | Rapid acceleration | | | | | | |
| | Rapid braking acceleration | | | | | | |
| | Intervehicular distance variation | | | | | ***.25 | *.16 |
| | Numbers of times of switching (ACC→Driver) | | | | | | |
| Driver driving control | Driving time | | | | | | |
| | Number of times of stepping on accelerator | | | | | | |
| | Number of times of stepping on the brake | | | | | | |
| | Rapid acceleration | | | | | | |
| | Rapid braking acceleration | | | | | | |
| | Intervehicular distance variation | | | | | | |
| | Numbers of times of switching (Driver→ACC) | | | | | | |

(continued)

Table 2. (continued)

| Factor | Independent variable | I | II | III | IV | V | VI |
|------------------------|---|--------|-------|--------|--------|---------|---------|
| Driver's driving style | Confidence in driving | | | | | | |
| | Passive mentality when driving | | | | | | |
| | Impatient driving tendency | | | *.21 | ***.26 | | ***-.26 |
| | Scrupulous driving tendency | *.21 | | | | ***-.36 | |
| | Signal-oriented preliminary preparing driving | | | | | | |
| | Driving as a status symbol | | | | | | ***.35 |
| | Unstable driving tendency | **-.28 | +-.17 | *-.18 | | | |
| | Natural worrier | ***.24 | | **-.27 | | **-.23 | |

In addition, 20 participants drove in six sections; therefore, we used 120 sampled data (20 participants × 6 sections) for the analyses. However, the driving style questionnaire was answered once for each participant. Hence, based on the assumption that the participants' driving styles were consistent, we used the average score of each element for each participant six times for the analyses.

Influence on Each Usability Element. First, the results of the analyses showed that regarding the driving control factor, the ACC and the ACC-driver driving control factors influenced the usability evaluation. However, the driver driving control factor did not influence the usability evaluation. The behaviors related to the ACC influenced the evaluation because the participants evaluated the ACC. Therefore, these results are considered to be valid. Moreover, the driver's driving style factor influenced the evaluation of all the usability elements. The influence on each usability element is given below.

Effectiveness. The participants who had scrupulous and stable driving tendencies and who naturally worried or were anxious about accidents evaluated the effectiveness of the ACC as high. These driving styles are considered to be related to safe driving. In fact, the ACC could keep a consistent distance between the car and the vehicle ahead and handle dangerous interruptions. The participants could drive safely with the ACC; therefore, the participants who had safe driving styles were assumed to provide a high evaluation of the effectiveness of the ACC. In the effectiveness evaluation, only the driver's driving style factor influenced the usability evaluation.

Efficiency. The participants who had a tendency of frequently stepping on the accelerator and had a short driving time in ACC-driver driving control evaluated the efficiency of the ACC as high. The frequent and short intervention in the ACC driving

control improved the evaluation; therefore, the ease of intervention could be considered to influence the efficiency. The participants who easily intervened in the ACC driving control provided a high evaluation of the efficiency.

Satisfaction. The participants who had impatient driving tendencies but were stable and were naturally anxious evaluated the ACC satisfaction level as high. The driving styles of drivers who were stable and naturally anxious also influenced the effectiveness. However, in evaluating the satisfaction, there was an influence of the impatient driving style. Impatient driving is related to actively overtaking the vehicle ahead and closing the distance with the vehicle ahead [10]. Therefore, the impatient driving style was considered to be related to the activeness of driving. In reality, when the vehicle ahead moved out of the range of the ACC radar sensor, such as when it changed the lanes, the ACC rapidly accelerated to the preset speed of 100 km/h at once. The participants could drive with safety and activeness using the ACC; therefore, the participants whose driving style was related to safety and activeness provided a high evaluation of the satisfaction level. In the evaluation of the ACC, safety as well as activeness is important.

Understandability. The participants evaluated the understandability as high because the ACC showed numerous accelerations (or decelerations) and a high value for the rapid braking acceleration. These would become high because the ACC adjusted the vehicle speed according to the surrounding environmental changes. Therefore, when the drivers apparently recognized the ACC driving control, the understandability is assumed to increase. In their interviews after the driving experience, the three participants clearly stated that when the ACC controlled the driving, they tended to make dangerous interruptions and rapid decelerations to understand the deceleration control of the ACC.

Discomfort. The discomfort was evaluated as low because of the small variance of the intervehicular distance in the ACC-driver driving control. The participants adjusted the vehicle position by intervening in the ACC driving; therefore, the variance of the intervehicular distance was considered to be small, and the discomfort in the ACC decreased. In other words, the discomfort associated with using the ACC decreased with the participants' direct adjustments of the vehicle positions.

Motivation. The independent variables that negatively influence the motivation tended to positively influence the other usability elements. The large variance in the intervehicular distance increased the efficiency and comfort, but it decreased the motivation. Moreover, the impatient driving style increased the satisfaction and understandability, but it decreased the motivation. These results indicated that the motivation for the drivers to drive by themselves would decrease because the ACC is useful.

3 Insight for Human-Automation System Interaction

First, the experimental results showed that the driver's driving style influenced the usability evaluation of the ACC. Rasmussen et al. [7] experimentally indicated that the driver's personality traits influence the usability evaluation of various kinds of systems.

They showed that users with different personality traits or different interests tended to use different strategies for the same system; consequently, they evaluated the effectiveness, efficiency, and satisfaction differently. Our results were consistent with those of the previous studies for the usability evaluation of the ACC, i.e., the driver's driving style influenced the usability evaluation. These results indicated that it is important to understand the results of the usability evaluation for automation systems by considering the personality traits. If there was no such consideration, the results of the usability evaluation could be misunderstood.

Next, the results of this study showed the influence on the usability evaluation, which could not be observed when cognitive artifacts were used. In using cognitive artifacts, such as computers, the users subjectively performed tasks, and the artifacts supported their activities [11]. In such situations, the ease of obtaining information inputs and outputs and the information processing speed and accuracy influence the usability evaluation of the artifacts [6].

However, in using automation systems, the systems subjectively conduct tasks, and the users monitor their activities and intervene in the system activities whenever necessary. In this study, the understandability results revealed that if the users recognize the system activities, the understandability of the system would increase. It is important for the users to clearly recognize the system activities in using automation systems. Furthermore, regarding the intervention in the system activities, the results about the evaluations of the efficiency and the discomfort revealed that the ease of use and the adjustability of the system activities influence the usability evaluation. Therefore, the efficiency and the discomfort would be enhanced if the users easily and directly adjusted the system activities even when the system does not behave according to the users' intentions. These results are specific for interactions with automation systems, and similar results could not be observed while using cognitive artifacts.

Finally, previous studies have pointed out the issue of disuse atrophy, which causes reduced human ability by overreliance on automation systems [12]. The results of the experiment indicated that because the ACC is useful, the motivation for the drivers to drive by themselves would decrease. The participants who consider driving as a status symbol evaluated the motivation as high; therefore, the participants who considered driving as a valuable activity are motivated to drive by themselves. These results indicate that the users were influenced by the types of activities that they perceived as valuable and were motivated to perform by themselves. Disuse atrophy might be prevented by letting the users subjectively perform such activities or by only supporting the activities using the automation systems.

4 Suggestions for ACC Design

In this section, we discuss the principle of the ACC design by considering the experimental results and the interviews. From the experimental results, we can see that the ease of intervention in the ACC driving control influenced the evaluation of efficiency. Also, the adjustability of the vehicle position influenced the evaluation of the discomfort. To enhance efficiency and reduce the discomfort, we could improve the

setup methods of the vehicle speed and the distance between the vehicles in the ACC driving control.

First, in regard to the setup methods of the ACC driving speed, the participants had to keep the control lever up or down until the speed was set as intended. In the interviews after the driving, some participants mentioned the setup method as “It was difficult to control the lever,” “The position of the lever was hard to find,” and “It takes too much time to control the lever.” To overcome the difficulty of setting up the vehicle speed, the lever could be replaced by a button or a switch so that the efficiency would be enhanced.

Next, for adjusting the intervehicular distances in the ACC driving control, the participants had to push a button to choose one of the three different distances. In the interviews, some participants made the following statements about the distance: “I would like to take the distance further” and “I would like to make a more minor adjustment.” The discomfort was assumed to decrease by making it possible to choose a distance from the more multiple distance levels and by making additional minor adjustments.

Moreover, the recognition of the ACC driving control influenced the understandability. Also, in the interviews, some participants described their understandability as follows: “In the beginning, I felt fear because I did not know how the ACC would behave, but I gradually understood the behavior and did not feel any fear” and “I understood the ACC behavior late in the driving.” To overcome this, we could display the state of the ACC driving control on the monitor in future; this would enable drivers to understand early the manner in which the ACC behaves. Also, drivers could experience using an ACC with a driving simulator; this would help them understand the ACC functioning before using it for real-life driving.

5 Conclusion

In this study, we investigated the determinative factors for the usability evaluation of an ACC. The results revealed that driving styles influenced all the usability elements. Also, the usability evaluation was not only influenced by the ACC driving control factor but also by the ACC-driver driving control factor. Based on our investigations into the human–automation system interactions, we have provided suggestions for improving the ACC design to enhance usability.

Acknowledgment. This work was supported by JSPS KAKENHI Grant Number JP16H02353 and by the Center of Innovation Program (Nagoya University COI: Mobility Innovation Center) from Japan Science and Technology Agency.

References

1. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* **39**(2), 230–253 (1997)
2. SAE International: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems. SAE J3016 (2014)

3. Maehigashi, A., Miwa, K., Kojima, K., Terai, H.: Development of a usability questionnaire for automation systems. In: Kurosu, M. (ed.) HCI 2016. LNCS, vol. 9731, pp. 340–349. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39510-4_32
4. ISO: Ergonomic requirements for office work with visual display terminals (VDTs) Part 11: Guidance on usability (1998). ISO 9241-11:1998
5. Neilsen, J.: Usability Engineering. Academic Press, Boston (1993)
6. Hornbæk, K.: Current practice in measuring usability: challenges to usability studies and research. *Int. J. Hum.-Comput. Stud.* **64**(2), 79–102 (2006)
7. Rasmussen, R., Christensen, A.S., Fjeldsted, T., Hertzum, M.: Selecting users for participation in IT projects: trading a representative sample for advocates and champions? *Interact. Comput.* **20**(2), 176–187 (2011)
8. Beggiano, M., Krems, J.F.: The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transp. Res. Part F* **18**, 47–57 (2013)
9. Beggiano, M.: Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transp. Res. Part F* **35**, 75–84 (2013)
10. Ishibashi, M., Okuwa, M., Doi, S., Akamatsu, M.: Indices for characterizing driving style and their relevance to car following behavior. In: Proceedings of SICE Annual Conference 2007, pp. 1132–1137 (2007)
11. Nornam, D.A.: Cognitive artifacts. In: Carroll, J.M. (ed.) *Designing Interaction: Psychology at the Human-Computer Interface*, pp. 17–38. Cambridge University Press, New York (1991)
12. Miwa, K., Terai, H.: Theoretical investigation on disuse atrophy resulting from computer support for cognitive tasks. In: Harris, D. (ed.) EPCE 2014. LNCS (LNAI), vol. 8532, pp. 244–254. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07515-0_25



Accent and Gender Bias in Perceptions of Interactive Voice Systems

Sabrina Moran^(✉), Ezekiel Skovron, Matthew Nare,
and Kim-Phuong L. Vu

California State University, Long Beach, CA 90840, USA
sabrina.n.moran@gmail.com, zeke@zekeskovron.com,
matthew.nare@gmail.com, kim.vu@csulb.edu

Abstract. Interactive Voice Systems (IVSs) are automated answering systems with pre-recorded menu options that are navigated by the user with keypresses and vocal responses. These systems are increasing in popularity with many companies because they facilitate the fielding of a large volume of callers, reduce personnel costs, improve efficiency, and increase callers' privacy; however, do people enjoy interacting with these systems? Individual characteristics, such as accent and gender, impact human interactions through biases, and it is likely some biases are maintained in interactions with voice systems. In the present study, participants were given scenarios and instructed to interact with the IVS to complete tasks. The accent and gender of the IVS were manipulated between scenarios, and perceptions of the IVS were gathered after each interaction in terms of pleasantness, likelihood of task completion, likelihood of recommending the system to a friend, and likelihood of using the system again. It was hypothesized that the results would align with findings relating to human interactions and bias, such that participants would prefer their native accent, and the female voice would receive more negative feedback. In contrast to the hypothesis, the Mexican accent was rated more pleasant, likely to be recommended to a friend, and likely to be used again than the American accent, regardless of the participants' native language. Overall, scenarios with positive outcomes were preferred over those with negative outcomes in all measures, and multiple interactions were found including the participant's first language and gender and the accent and gender of the IVS.

Keywords: Automation and autonomous systems
Computer-mediated communication · Designing for pleasure of use
Display design · Human centered design · Human factors/system integration
Interactive Voice Systems · Social bias

1 Introduction

As our culture continues to advance technologically, it is natural that we adapt our everyday lives accordingly. In the last decade, technology use has profoundly increased, with many people routinely relying on technological devices every day to connect them to friends and family, supply navigational instruction, facilitate immediate access to information, and much more. Mainstream use of technology puts

pressure on designers to create high-performing and coherent systems that can accommodate multiple demographic groups (e.g. different ages, genders, ethnicities). One family of systems that has recently become very popular is the interactive voice system (IVS). These automated answering systems include prerecorded menu options that are relayed to the caller, whom can navigate through a vocalized menu using key presses or vocal responses to complete actions or be directed to the correct customer service personnel.

IVSs can serve large quantities of people more efficiently and conveniently than a large pool of employed personnel, reducing the company costs and increasing customer privacy when entering personal information [1]. Everyday consumers often interact with these systems without thought, as they have already been adopted by many large companies, such as pharmacies, insurance agencies, and banks. As IVSs become more widely used, it is imperative that they promote effective, enjoyable use to increase their acceptance by the public. Whether a user is following the directions of a Global Positioning System (GPS) to an unfamiliar location or making a doctor's appointment, it is important that upon finishing the interaction, the user feels like s/he has successfully completed the intended task.

Prior to the introduction of IVSs, humans relied solely on other humans to accomplish many of the tasks that IVSs are now able to complete. Although IVSs are not designed to be a human, they were made with the intention of mimicking human interactions to some degree. Therefore, it can be assumed that many of the empirical findings relating to interactions between individuals can also be applied to interactions with IVSs. In a human interaction study, Mayer et al. [8] found that trust plays an important role in reliance between individuals, and it is likely that this finding applies to IVSs as well. Minimal research has investigated the factors underlying trust in IVSs specifically, but research investigating human interactions has provided insight into possible factors that impact perceived trustworthiness. Gluszek and Dovidio [4] found that factors, such as gender and ethnicity, influence one's trust in other individuals. It is also possible that ingroup and outgroup bias, from the formation of social groups, may influence trust and human behavior. In-group favoritism is a phenomenon that tells us there is a tendency, under certain conditions, to favor members of groups to which an individual belongs to (ingroups) over groups that an individual does not belong to (outgroups). For example, people tend to judge accents most similar to their own as more trustworthy and favorable [1, 11].

Accent bias has been found in children as young as five months old with a preference for individuals of their native accent shown through longer visual fixations [5]. When given a choice, young children in the US and France prefer to be friends with peers with their native accent [5]. Furthermore, a study by Wang et al. [11] investigated how the accent of a customer service employee impacts his or her performance ratings. They found that the Indian customer service representative received lower performance ratings than that of the American or British accent representative in scenarios with identical customer service interactions. Gluszek and Dovidio [4] also found a prejudice present against nonnative accents, such that individuals with nonnative accents were considered less intelligent, less loyal, more negative, and less competent. Moreover, regardless of the actual language competency of an individual with a nonnative accent, they were perceived as not speaking the language as fluently as someone with a native

accent [4]. Supporting these findings, Verberne et al. [10] concluded that an increase in similarity between two humans, or a human and a virtual agent, will likely lead to an increase in trust in the system.

One of the few influential studies relating to IVS interactions was completed by Large and Burnett [7], and the study investigated how participants rated the trustworthiness of an interactive voice GPS in a driving simulator. Researchers compared the participant's perceptions of a default male voice and a voice resembling "Snoop Dogg," a rapper from Southern California. While participants rated the default male voice as more trustworthy than "Snoop Dogg," this perceived trust did not impact their actions when interacting with the system [7]. When the participants were presented with a conflicting situation in which they had to decide whether to follow the instructions of a road sign or the GPS, their decisions were not consistent with their reported bias towards the more trustworthy GPS. These findings support the notion that the attitude and voice of the system impacts trust, with clearer, more intellectual voices receiving higher trust ratings from participants; however, the bias toward or against different accents is not always significant enough to influence one's actions. While these findings may seem inconclusive, it is important to keep in mind that only two voices were being compared by Large and Burnett, and both voices were versions of the American vernacular. As IVSs increase in popularity, it is essential that society is comfortable with its normalized use, and research on accent trust bias can potentially make this normalization easier for the average user through the revelation of population trends.

The potential for gender bias is another aspect that should be considered. Ko et al. [6] suggested that it is common to make gender-based assumptions contingent on vocal cues. Ko et al. found that the gender of the IVS influences how the users perceive it and how much trust they place in the system. Additionally, Werner and LaRussa [12] found that men were perceived to be more dominant and forceful, while women were perceived to be more warm and sensitive. Overall, the female voices received significantly more negative comments than the male voices, even with an identical script, revealing a bias against the female voices.

In human interactions, it has been demonstrated that face-to-face interviews in the medical environment regarding personal habits produce gender biases, with the female interviewer obtaining more information from the patient compared to the male interviewer [9]. In the medical field, it is very important to ensure that patients are comfortable when being asked about sensitive information, and it is possible that the gender of the interviewer, human or IVS, may impact the quality of the information gathered. While this finding could be extremely useful in the medical field, it is possible that the potential increase in comfort with the female interviewer may only apply to human interactions. Evans and Kortum [3] investigated what voice personalities promoted trust from the user in medical IVSs. Participants heard differing male and female voices and rated the extent to which they liked and trusted the voice. Evans and Kortum did not find a significant difference in the rated trust or liking of the different gendered voices. A few years later, Edwards and Kortum [2] investigated how the voice of an IVS influenced its perceived usability. They utilized upbeat, professional, and sympathetic voice personalities for both the male and female voices, and found that, overall, the male voice received higher subjective ratings of usability among

participants compared to the female voice. However, the gender of the voice did not impact how participants viewed the system's learnability.

The present study aims to extend our understanding of how findings relating to accent and gender biases in human interactions carryover into IVS interactions. The goal of the study is to investigate the influences that the accent and gender of an IVS have on the user's perceptions of the system, as well as, whether the gender and first language of the user would influence these effects. These findings will potentially assist the creation of future IVSs, and as IVSs continue to increase in popularity among businesses, it is important to ensure that they are functioning at the highest level possible, with the greatest user experience, to promote system efficiency and acceptance. It was hypothesized that many of the results found in the past human interaction research will be maintained in the IVS interactions. Thus, we expected to find that the female voice would receive lower ratings than the male voice, and users would rate voices of their native accent higher than their nonnative accent. Additionally, it was hypothesized that scenarios with negative outcomes would be rated lower when the IVS speaks with a nonnative accent than with a native accent or with a female voice rather than a male voice.

2 Method

2.1 Participants

A total of thirty-three college students (10 males; 23 females), over the age of 18, participated in this study. Of the participants, 19 were native English speakers, 8 were native Spanish speakers, and 6 were native speakers of a different language. All participants were undergraduate students recruited at California State University, Long Beach through the Introduction to Psychology (PSY 100) subject pool.

2.2 Materials

The study consisted of 25 task-based scenarios created by the researchers. The 25 tasks were split evenly between five different contexts: GPS Navigation, Doctors Appointment Service, Pharmacy Medication Refill, Grocery Delivery Service, and Delivery Company. All scenarios necessitated the participant to interact with the IVS to complete a task, and in each context, one of the five interactions resulted in a negative outcome, in which the participant was informed that their goal was not accomplished by the system. The voices used for the interactions varied between two accents (American and Mexican) and two genders (Male and Female), and the tasks and voice conditions were randomly intermixed and counterbalanced between participants.

The voices for the IVS were recorded using Macintosh's "Text to Speech" function, and the scenarios were programmed using Visual Basic in Visual Studio, allowing for the entire study to be self-contained within a single interactive program. The program controls the interactivity of the windows, the counterbalanced ordering of the tasks, and ensures all within-subject conditions were manipulated to a predefined counterbalance matrix. All data was internally gathered by the program and stored in a text file.

A single Windows desktop computer was used for all participants with two forward facing speakers and a keyboard and mouse for navigation and actions. A sample screenshot from one interaction is shown in Fig. 1.

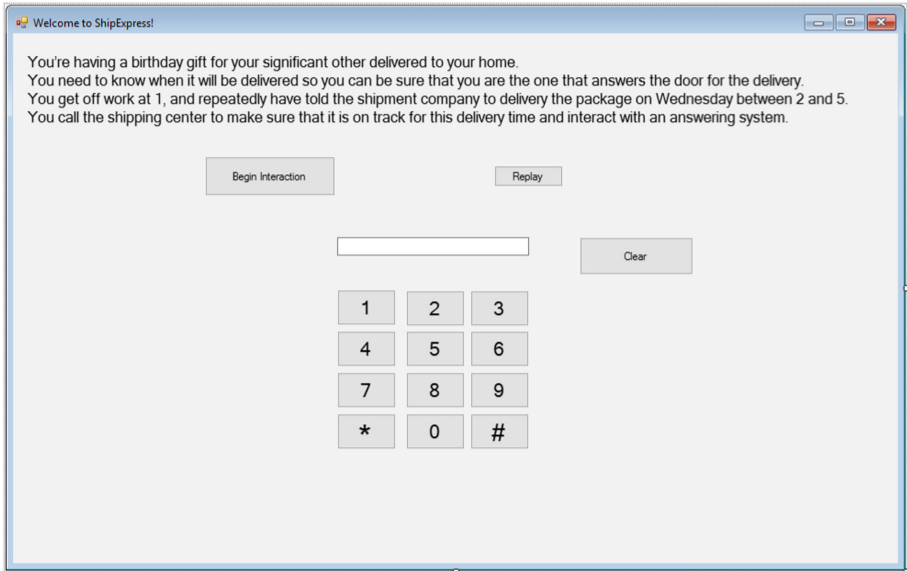


Fig. 1. A sample screenshot of one of the scenarios used in the present study.

2.3 Protocol

Upon arrival, participants were given a short description of the study at hand, including a definition and examples of an IVS. Then, participants were asked to read and sign an informed consent form prior to filling out a demographic questionnaire. Following this, participants were given instructions on how to interact with the system and were asked to begin the first interaction. A counterbalance matrix was used to control for the order of the scenarios to ensure no learning effects occurred and to control the combinations of the gender and accent of the IVS for each of the scenarios. The accent and gender of the IVS was counterbalanced to ensure that each participant experienced every accent and gender combination as equally as possible and in varying orders. After every interaction, participants were presented with a questionnaire that displayed four Likert-scale questions asking the participant to rate the interaction in terms of pleasantness (pleasantness measure), likelihood of task completion (completion measure), likelihood of recommending to a friend (recommend measure), and likelihood of using this system again for a future task (use again measure). Questionnaire responses were internally stored by the program.

3 Results

A mixed 2 (Participant’s Gender) × 2 (Participants’ First Language) × 2 (Gender of IVS) × 2 (Accent of IVS) × 2 (Scenario Outcome) ANOVA was completed for each of the four dependent measures. The between-subject factors were the participants’ gender and first language, and the within-subject factors were the gender of the IVS, the accent of the IVS, and the outcome of the interaction. A summary of the significant effects can be found in Table 1.

Table 1. Summary of significant effects found in Mixed ANOVA.

| Effect | Pleasant | Completion | Recommend | Use again |
|--|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| IVS Accent | $F(1,27) = 8.90$ $p = .006$ | | $F(1,27) = 21.37$ $p < .001$ | $F(1,27) = 21.97$ $p < .001$ |
| Outcome | $F(1,27) = 39.59$ $p < .001$ | $F(1,27) = 64.97$ $p < .001$ | $F(1,27) = 31.99$ $p < .001$ | $F(1,27) = 23.96$ $p < .001$ |
| IVS Gender × P Gender | $F(1,27) = 6.21$ $p = .019$ | | | |
| IVS Gender × P First Language | $F(2,27) = 5.95$ $p = .007$ | | | $F(2,27) = 4.87$ $p = .016$ |
| IVS Gender × Outcome | $F(1,27) = 4.87$ $p = .036$ | | $F(1,27) = 7.36$ $p = .011$ | $F(1,27) = 9.92$ $p = .004$ |
| IVS Accent × Outcome | $F(1,27) = 21.07$ $p < .001$ | $F(1,27) = 13.54$ $p = .001$ | $F(1,27) = 21.39$ $p < .001$ | $F(1,27) = 18.27$ $p < .001$ |
| IVS Gender × Outcome × P Gender | $F(1,27) = 7.39$ $p = .011$ | | | |
| IVS Gender × Outcome × P First Language | $F(2,27) = 3.66$ $p = .039$ | | | $F(2,27) = 3.77$ $p = .036$ |
| IVS Accent × Outcome × P Gender × P First Language | | $F(2,27) = 4.24$ $p = .025$ | | |

There were two significant main effects and seven significant interactions. As can be seen in Table 1, the accent of the IVS’s voice significantly impacted three of the measures: the perceived pleasantness, likelihood of recommending to a friend, and the likelihood of using the IVS again. For these measures, the Mexican accent was rated better than the American accent (see Fig. 2).

Visible in Table 1, a main effect of outcome was also found in all four measures. Unsurprisingly, scenarios with positive outcomes were rated better than scenarios with negative outcomes in perceived pleasantness, likelihood of task completion, likelihood of recommending to a friend, and likelihood of using the system again (see Fig. 3).

Seven significant interactions were also found during analysis. The only interaction associated with the native accent bias was the 4-way interaction of the accent of the IVS, the first language of the participant, the outcome of the scenario, and the participant’s gender for the completion measure (see Table 1). We found that male

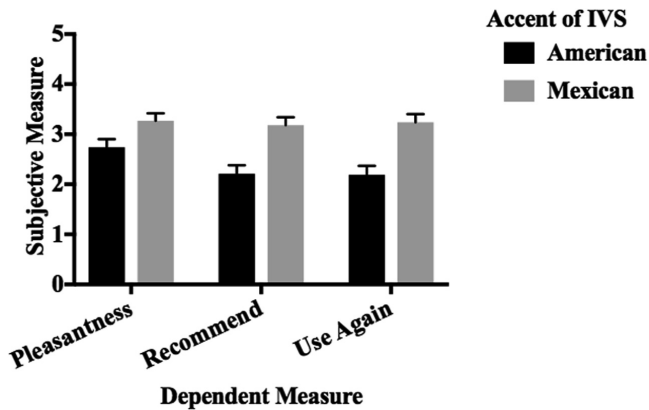


Fig. 2. The main effect of the accent of the IVS found in the perceived pleasantness, likelihood of recommending to a friend, and likelihood of using again. All scores are based on Likert-scale responses, with higher values equating to higher agreeability with the measure.

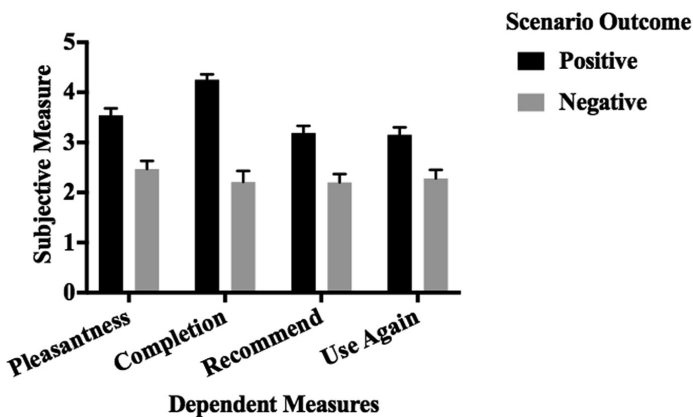


Fig. 3. The main effect of the outcome of the scenario found in the perceived pleasantness, likelihood of task completion, likelihood of recommending to a friend, and the likelihood of using again. All scores are based on Likert-scale responses, with higher values equating to higher agreeability with the measure.

participants who were native English speakers rated the IVS Mexican accent to be better than the IVS American accent regardless of outcome. Male participants who were native Spanish speakers or native speakers of another language besides English or Spanish also rated the Mexican IVS accent as better for scenarios with positive outcomes. However, for scenarios with negative outcomes, male participants rated the IVS with an American accent as more likely to have completed their task than the IVS with a Mexican Accent (see Fig. 3). Female participants who were native English speakers or native speakers of another language besides English or Spanish rated the IVS Mexican accent to be better than the IVS American accent regardless of outcome.

Female participants who were native Spanish speakers also rated the Mexican IVS accent as better for scenarios with positive outcomes; however, they showed no preference for either IVS accent in scenarios with negative outcomes (see Fig. 4).

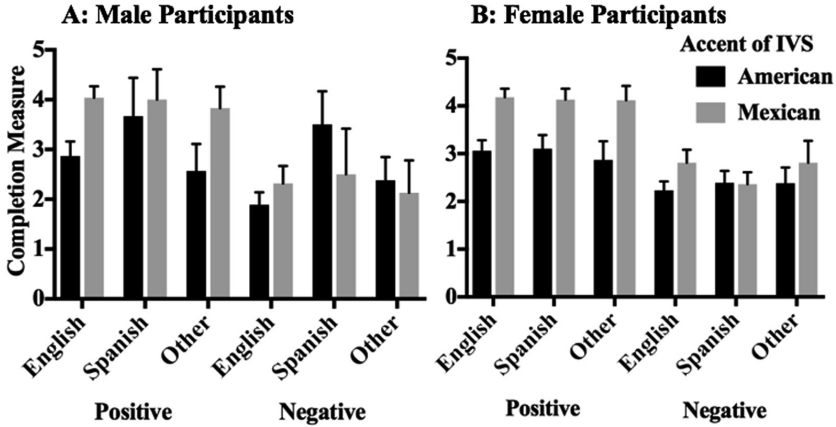


Fig. 4. The interaction between the accent of the IVS, the outcome of the scenario, the participant’s gender, and the participant’s first language. Relevant data is shown for the (A) male participants and the (B) female participants. All scores are based on Likert-scale responses, with higher values equating to higher agreeability with the measure.

There was a two-way interaction between gender of the IVS’s voice and the outcome of the scenario (see Table 1). For all three of the measures of pleasantness, recommend to a friend, and likelihood of using the system again, the male voice was preferred to the female voice in scenarios with positive outcomes, but the reverse was evident in scenarios with negative outcomes (see Fig. 5).

Another significant interaction between the accent of the IVS’s voice and the outcome of the interaction was found in all four of the dependent measures (see Table 1). The Mexican accent was rated higher in the scenarios with a positive outcome for all measures. In scenarios with a negative outcome, participants still rated the Mexican accent higher than the American accent in the perceived pleasantness, likelihood of recommending to a friend, and the use again measure; however, in the likelihood of task completion measure, participants rated the American accent higher than the Mexican accent (see Fig. 6).

The remaining interactions all involved the effects of gender. An interaction was found between the gender of the IVS and the gender of the participant in the pleasantness measure (see Table 1). Male participants rated the female voice more pleasant, and female participants rated the male voice more pleasant (see Fig. 7).

Another significant interaction was found, in the pleasantness and the use again measures, between the first language of the participant and the gender of the IVS’s

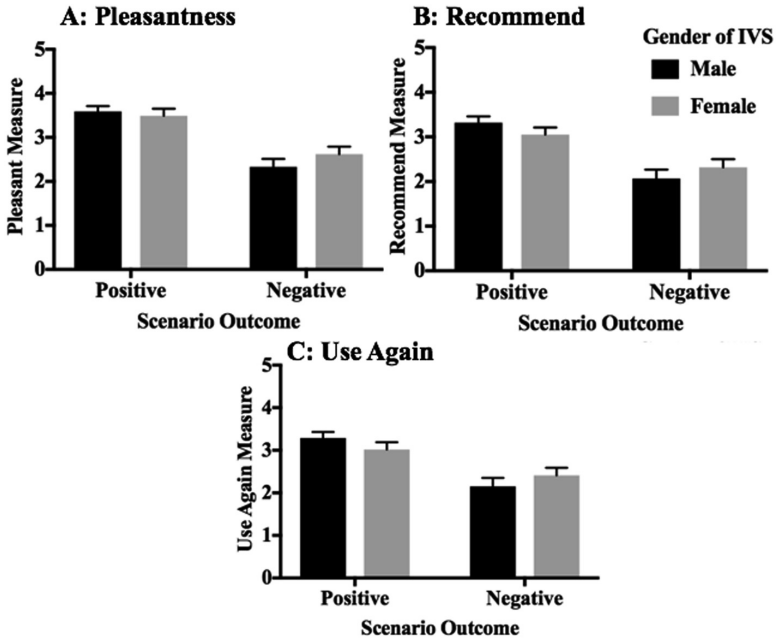


Fig. 5. The interaction between the gender of the IVS and the outcome of the scenario in (A) perceived pleasantness, (B) likelihood of recommending to a friend, and (C) the likelihood of using again measure. All scores are based on Likert-scale responses, with higher values equating to higher agreeability with the measure.

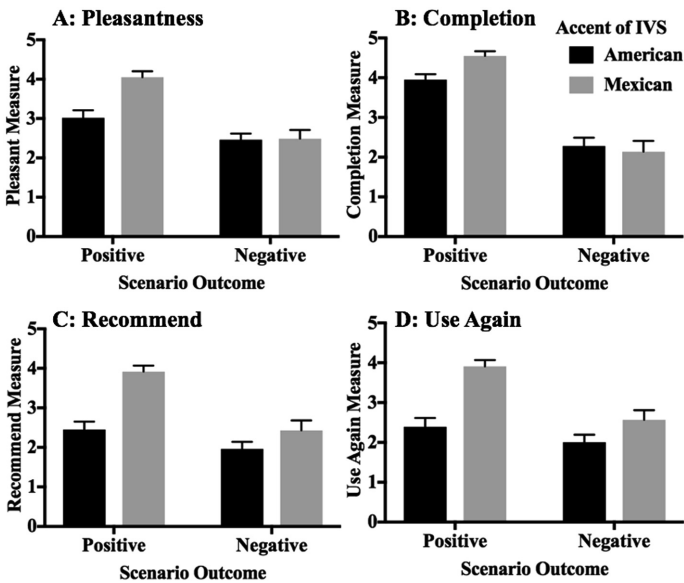


Fig. 6. The interaction between the accent of the IVS and the outcome of the scenario in the (A) perceived pleasantness, (B) likelihood of task completion, (C) likelihood of recommending to a friend, and (D) likelihood of using again. All scores are based on Likert-scale responses, with higher values equating to higher agreeability with the measure.

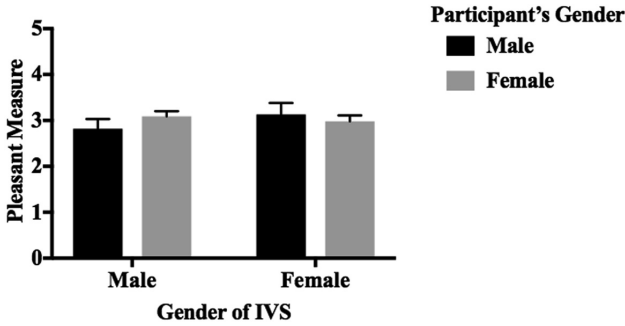


Fig. 7. The interaction between the gender of the IVS and the gender of the participants in the perceived pleasantness of the interaction. All scores are based on Likert-scale responses, with higher values equating to higher agreeability with the measure.

voice (see Table 1). For both measures, native English speakers rated the male voice higher the female voice, native Spanish speakers rated the female voice higher than the male voice, and native speakers of a language other than English or Spanish showed little gender preferences (see Fig. 8).

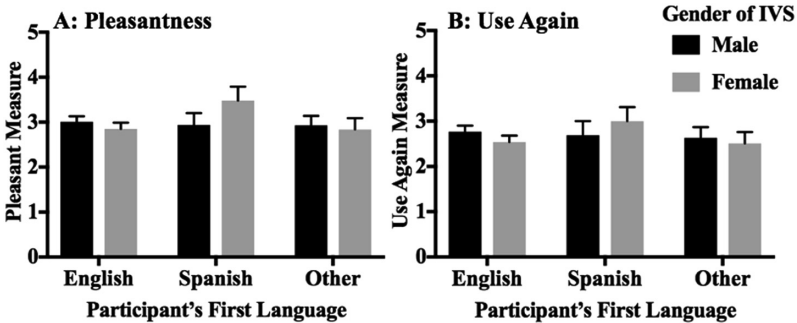


Fig. 8. The interaction between the participant's first language and the gender of the IVS in the (A) perceived pleasantness and (B) likelihood of using again. All scores are based on Likert-scale responses, with higher values equating to higher agreeability with the measure.

A fifth interaction was found between the gender of the IVS, the outcome of the scenario, and the participant's gender in the rated pleasantness of the interaction (see Table 1). Male participants rated the male voice more pleasant in the positive scenarios and the female voice more pleasant in the negative scenarios, and the female participants rated the male voice more pleasant in both the positive and negative scenarios (see Fig. 9).

The sixth significant interaction, between the IVS gender, the outcome of the scenario, and the first language of the participant, was found in the pleasantness and use again measures (see Table 1). In both measures and scenario outcomes, the participants

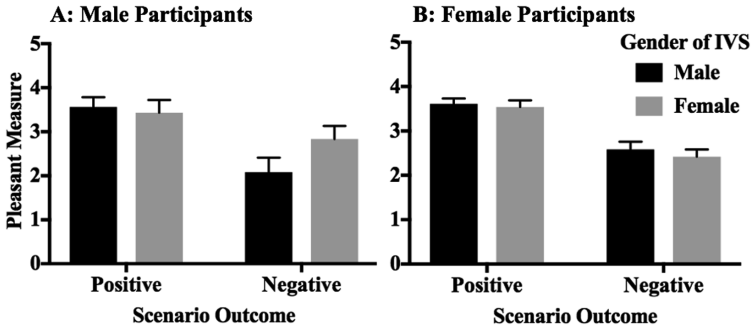


Fig. 9. The interaction between the gender of the IVS, the outcome of the scenario, and the participant’s gender. Relevant data is shown for (A) male participants and (B) female participants. All scores are based on Likert-scale responses, with higher values equating to higher agreeability with the measure.

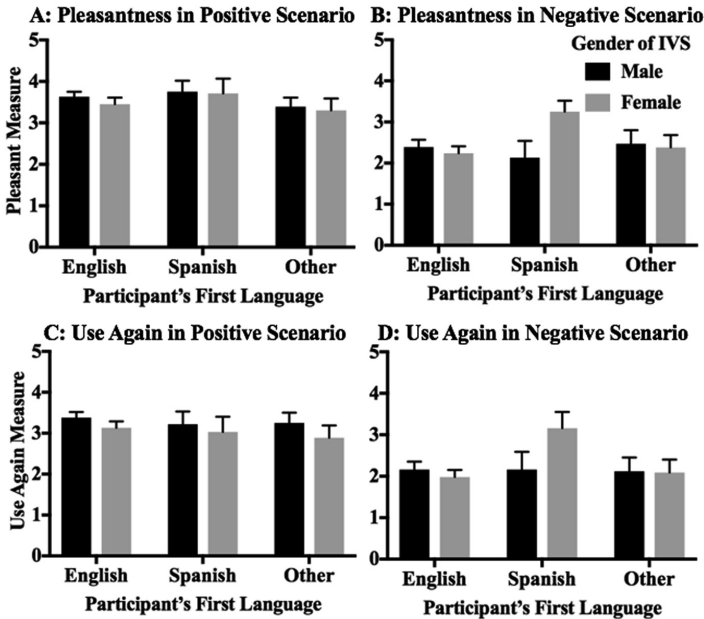


Fig. 10. The interaction between the gender of the IVS, the participant’s first language, and the outcome of the scenario in the pleasantness and use again measures. Relevant data for the pleasantness measure is displayed for (A) positive and (B) negative scenarios, and relevant data for the use again measure is displayed for the (C) positive and (D) negative scenarios. All scores are based on Likert-scale responses, with higher values equating to higher agreeability with the measure.

that were native English speakers and native to another language rated the male voice higher than the female voice. The participants that were native Spanish speakers rated the male voice higher than the female voice in the pleasantness and use again measures in the positive scenarios, but in the negative scenarios, the native Spanish speakers rated the female voice higher in both measures (see Fig. 10).

4 Discussion

The purpose of the present study was to investigate whether accent bias and gender bias tendencies recognized in face-to-face interactions are maintained in interactions with IVSs. Analysis revealed that the accent of the IVS impacts the perceived pleasantness, likelihood of recommending the system to a friend, and the likelihood of using the system again, such that the Mexican accent was rated higher in all three measures. This is contrary to our hypothesis and past accent bias findings due to a majority of our participants being native English speakers; therefore, a preference for the Mexican accent is evidence of a preference for a nonnative accent.

Little evidence was revealed in the present study supporting accent bias. The only significant interaction with findings relating to native accent bias revealed an interaction in the likelihood of task completion measure with the first language of the participant, the accent of the IVS, the outcome of the scenario and the gender of the participant. Both genders rated the IVS with a Mexican accent more likely to complete their task in all positive scenarios, but in the negative scenarios, the participant's gender and native language impacted the scores given to the IVS accents. Male participants that were native Spanish speakers or native to another language rated the IVS with an American accent more likely to complete their task than the Mexican accent in the negative scenarios. Female participants that were native English speakers or native to another language maintained a bias towards the Mexican accent in the negative scenarios, and female participants that were native Spanish speakers did not show an accent preference. This interaction provides evidence of a preference for a nonnative accent with the understanding that the gender of the participant may be a factor.

Many interactions were found relating to gender bias. When the data was inspected based on the participant's gender, the males rated the female voice more pleasant, and the females rated the male voice more pleasant. This gender bias against one's own gender is a phenomenon that should be studied further with a larger, evenly distributed sample. Results also revealed that the first language of an individual may impact their gender preference. Native English speakers and participants native to another language rated the male IVS voice more pleasant and likely to be used again than the female voice; however, the native Spanish speakers rated inversely, with the female voice rated more pleasant and likely to be used again. These outcomes support the notion that gender bias may depend on the outcome of the scenario and the native language of the participant. Furthermore, in positive scenarios, the male IVS voice was rated more pleasant, likely to be recommended, and likely to be used again than the female voice; however, in the negative scenarios, the female voice received higher ratings in these categories.

The findings from positive scenarios somewhat supports the second hypothesis that female voice would receive poorer ratings, but this effect was not maintained in negative scenarios. The final hypothesis of a preference for native accents and male voices in scenarios with a negative outcome was not supported. Contrary to the hypotheses, in the scenarios with a negative outcome, the female voice and the nonnative accent were rated higher than the male voice and the native accent, respectively, in all measures except the completion measure.

This study was completed as a semester-long directed research project; therefore, time and resources were limited. Some mentionable limitations of this study include the sample size, uneven distribution among between-subject groups (participant's gender and first language), and the quality of the voice recordings. It is possible that if the sample size for the study had been larger, more definitive results may have been found. Additionally, statistical power would have been improved if the participants had been evenly distributed between the gender groups and first language groups. In addition, participants for this study were gathered from a Psychology 100 subject pool; therefore, the researchers did not have the ability to select participants or control for between-subject factors. Finally, the voices used for the study were recording from a Macintosh "Text to Speech" function; therefore, voice selections for the American and Mexican accents were limited and the quality of the accents may not be representative.

As we continue to find new ways to use technology in various contexts, like IVSS, it will continue to be important to complete studies investigating how to tailor the technology to promote positive user experience and efficient interactions. The current findings provide an initial investigation, and more research is needed to supply conclusions relating to what aspects of an IVS impact the user's perceptions of the interactions. The results of the present study act as a clear sign that we cannot assume all human interaction findings will translate into the computer-based interactions. Future research should investigate deeper into whether the context of the interaction impacts the preferred characteristics of the IVS's voice, and more research should investigate accent bias on a larger scale.

Acknowledgements. We would like to extend gratitude to Ryan Fitz and Laura Yorba for participating in the research group that brainstormed the initial idea for this study. Additionally, we would like to thank the Center for Usability in Design and Accessibility (CUDA) lab at California State University, Long Beach for giving us the opportunity and resources for the execution of this study.

References

1. Dulude, L.: Automated telephone answering systems and aging. *Behav. Inf. Technol.* **21**(3), 171–184 (2002). <https://doi.org/10.1080/0144929021000013482>
2. Edwards, R., Kortum, P.: He says, she says: does voice affect usability? *Proc. Hum. Fact. Ergon. Soc.* **56**(1), 1486–1490 (2012). <https://doi.org/10.1177/1071181312561295>
3. Evans, R.E., Kortum, P.: Voice personalities inducing trust and satisfaction in a medical interactive voice response system. *Proc. Hum. Fact. Ergon. Soc.* **53**(18), 1456–1460 (2009). <https://doi.org/10.1037/e578522012-061>

4. Gluszek, A., Dovidio, J.: The way they speak: a social psychological perspective on the stigma of nonnative accents in communication. *Pers. Soc. Psychol. Rev.* **14**(2), 214–237 (2010). <https://doi.org/10.1177/1088868309359288>
5. Kinzler, K.D., Shutts, K., DeJesus, J., Spelke, E.S.: Accent trumps race in guiding children's social preferences. *Soc. Cogn.* **27**, 623–634 (2009)
6. Ko, J.K., Judd, C.M., Blair, I.: What the voice reveals: within- and between category stereotyping on the basis of voice. *Pers. Soc. Psychol. Bull.* **32**(6), 806–819 (2006). <https://doi.org/10.1177/0146167206286627>
7. Large, D., Burnett, G.: The effect of different navigation voices on trust and attention while using in-vehicle navigation systems. *J. Saf. Res.* **49**, 69–75 (2014). <https://doi.org/10.1016/j.jsr.2014.02.009>
8. Mayer, R., Davis, J., Schoorman, F.: An integrative model of organizational trust. *Acad. Manag. Rev.* **20**(3), 709–734 (1995). <https://doi.org/10.2307/258792>
9. Pollner, M.: The effects of interviewer gender in mental health interviews. *J. Nerv. Ment. Dis.* **186**, 369–373 (1998)
10. Verberne, F., Ham, J., Midden, C.: Trusting a virtual driver that looks, acts, and thinks like you. *Hum. Factors* **57**(5), 895–909 (2015). <https://doi.org/10.1177/0018720815580749>
11. Wang, Z., Arndt, A., Singh, S., Biernat, M.: The impact of accent stereotypes on service outcomes and its boundary conditions. *NA - Adv. Consum. Res.* **36**, 940–941 (2009)
12. Werner, P., LaRussa, G.: Persistence and change in sex-role stereotypes. *Sex Roles* **12**(9–10), 1089–1100 (1985)



Tangible User Interface

Elias Shamilov¹, Nirit Gavish², Hagit Krisher^{3(✉)}, and Eran Horesh²

¹ Intactio, 2029600 Eshchar, Israel
Elias@intactio.com

² ORT Braude Academic College, Karmiel, Israel

³ Technion, Israel Institute of Technology, Haifa, Israel
hagitkrisher1@gmail.com

Abstract. We introduce here a tangible user interface implemented by passing objects between a human and a computer.

This interface deals with realization and virtualization: when we realize an object, it disappears from the screen and we get a real object out of the screen, as it were. Virtualization is the opposite action: we pass real object into the user interface, it disappears, and we get a virtual object on the screen.

Our experiment included a grocery shop game with a coin machine as an input-output device. The virtual money that was earned in the game was realized as metal coins. The metal coins paid into the user interface were represented virtually on the computer screen.

The experiment included two ages groups: adults and children 6–8 years old. Reference groups played the same game without tangible components, i.e. all the money in the game was virtual (a common implementation in the game industry).

Adults bought less products in the version with tangible interface, and the decision period was longer in that case.

The children's results were completely different: they bought many more products when playing the game with the tangible interface (about 3 times more). Children perceived the game with tangible interface as a totally different game: intuitively understandable and more engaging.

These results accord with cognitive theory, and we are look forward to further research studying the effectiveness of the use of tangible user interfaces.

Keywords: Tangible · Augmented reality · Mixed reality · Augmented virtuality
Internet of things · Teleportation · User interface

1 Introduction

Human-computer interaction has come a long way since the first computers: from punch cards in the first computers, through keyboard-only terminals and, later, mouse-based graphical interface, and on to various touch screens, voice control and gesture interfaces.

Human-computer interfaces have become more intuitive: if punched cards could be read only by experienced programmer, the touch screen concept is understandable for three years old kids.

Another aspect of interface development is the blurring of differences between virtual reality and the real world. The phrase “Augmented Reality” is used commonly and not only in science fiction books.

User interfaces have become much more accessible and similar to concrete reality. A human interacts with a computer in the same way as he does with other humans and gets intuitive, understandable feedback. Those technologies are not yet well-developed, but we can already see the direction in which user interfaces are developing: humans can speak to a computer and a computer can speak back; a human can see a virtual world (or augmented world) around him and move virtual items using gestures, he can get physical feedback from the computer – for example a device-costume can hit the user.

Virtual reality interfaces allow humans to sense the virtual world just as they sense the physical world. The only thing that prevents humans from complete immersion in a virtual world is the fact that the items that belong to the virtual world remain there and only there. A human can move them by more-or-less intuitive gestures from one virtual place to another, but it is impossible to move items from the virtual world into the real world. Virtual artifacts remain virtual, we can't touch them, and they disappear when the virtual world is turned off.

We challenge this restriction. We define a tangible interface and blur the last border between real and virtual worlds. We allow the user to turn virtual objects into physical objects and vice versa. To use HCI terminology: the human-computer interaction now includes passing objects between human and computer.

2 Background

2.1 The Tangible Interface – Description

The “objects passing between human and computer” here should be understood literally.

In one direction, the user can drag a virtual (drawn) object out of the screen in the same way he can drag it on the top of a touch screen. In this case the item will “get out” of the screen. (a virtual, drawn item will disappear from the screen and a real, physical item will appear outside the screen) – the item will be realized.

Or, in the opposite direction, the user can throw an item into the screen. The real item will disappear (or get out of the user's range) and a virtual item will be shown on the screen. The item will be virtualized.

To enable this activity, we build a new input-output device called Realizer/Virtualizer, that recognizes and consumes physical items in the case of virtualization and emits physical items in the case of realization.

2.2 Theoretical Background

The processing of the information that we receive from the physical world involves all the senses. Cognitive research shows that visual data is much more understandable than audio data for most people, but the tactile sensations are more effective when dealing with little children. They study their surrounding by touching things, and their ability to understand the meaning of virtual artifacts is limited. For this reason, we expect that the

tangible interface will involve the children in the experience and increase their understanding, creating an enhanced user experience.

In the case of adults, the difference seems less dramatic, but still we expect the tangible interface to improve their attitude and the perception of the game.

2.3 Classification

The concept of items' transportation between real and virtual worlds is new and this technology doesn't fall into any known category.

At first one might be inclined to define it as "augmented reality". Indeed, the system includes virtual and physical parts. However, the phrase "augmented reality" refers to mixed real and virtual visualization of the world. Here the technology is different; it includes real and virtual parts, but the real parts stay real and the virtual parts stay virtual. Both worlds coexist, and items can easily and intuitively be transported between them. Although the terminology of "augmented reality" is not limited to visual effects only, the usage of such terminology here misses the target. After all, this technology can be used together with augmented reality technology or without it.

In the taxonomy of Benford (2) (see Fig. 1) the place of tangible interface even is not uniquely defined: there is no place for an interface that is virtual and real at the same time.

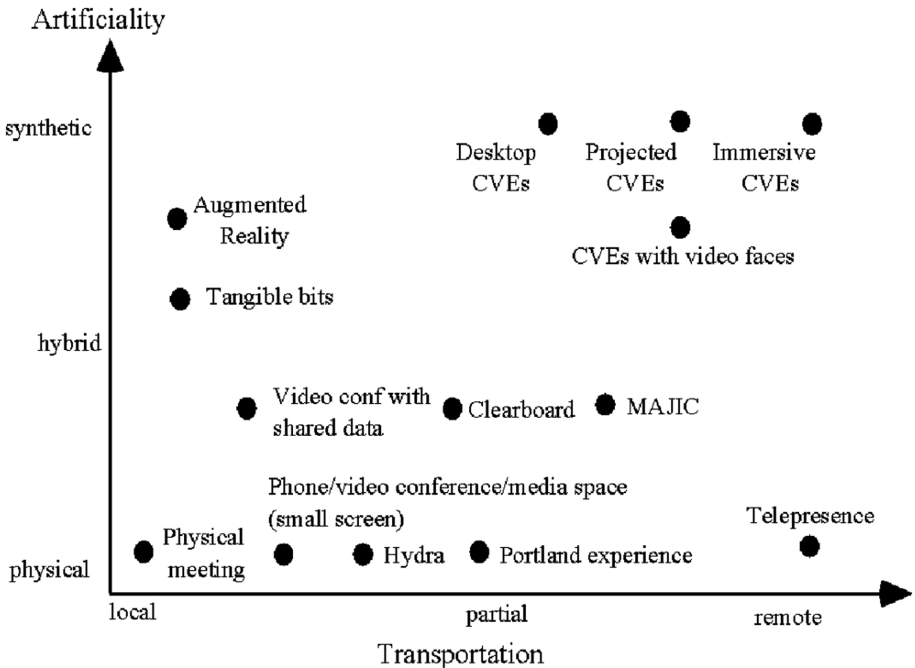


Fig. 1. Benford's taxonomy of augmented realities

One might also attempt to associate this technology with the “internet of things”. IoT augments the capabilities of real world objects by connecting them to the Internet, but this does not create objects which transition from physical to virtual reality. Physical items are recognized by computers and affect their behavior, but the real stays real and the virtual stays virtual. Further, tangible user interface technology can be used without, with or in addition to IoT technology.

This technology describes a new human-computer interface type, that has not yet been the object of serious research. Most references to such technology are in science fiction books, not a source we can study productively.

3 Experiment Description

3.1 System and Setup

This paper describes the evaluation of a system with a tangible interface implemented as a computer game. The game was played on regular 2D screen (no use of virtual or augmented reality by VR glasses), no voice command or gesture interfaces were used. The system that was used in the experiment used standard interfaces along with the tangible interface described here.

The tangible items that were used in the experimental game were metal coins of 27.5 mm diameter, represented virtually as a gold coin image of 39 pixels width. An input-output device was used: each time a user ordered the removal of a coin out from the screen, the virtual embodiment of the coin disappeared, and the physical coin fell out of the coin machine. Each time a user threw a coin into the coin machine, the virtual coin appeared on the screen.

In the game, the player is a seller in a grocery shop, customers come and ask to buy products, the player gives them the products and the customer takes the products and pays. Figure 2 shows an example of a screenshot when the customer arrives.

When the supplier arrives, he offers products for sale. In this scenario, the player should decide if he buys products and how many (he can choose not to buy). If the player chooses to buy, he must pay for the product or products.

A control group played the same game with only one difference: earned coins were stored in a virtual wallet – a gold coin image and a number drawn near it, the common representation of virtual wallets in computer games.

The experiment measured the change in the perception of virtual world when using the tangible interface. Because children and adults perceive the worlds differently, the experiment was done on two different groups:

1. Adults (in their 20 s).
2. Children (ages 6–8).



Fig. 2. The experimental system – game screenshot

3.2 Experiment Flow

As mentioned earlier, the experiment included two test groups: adults and children. In each part thirty participants used the tangible interface version and thirty used the virtual version (Figs. 3 and 4).

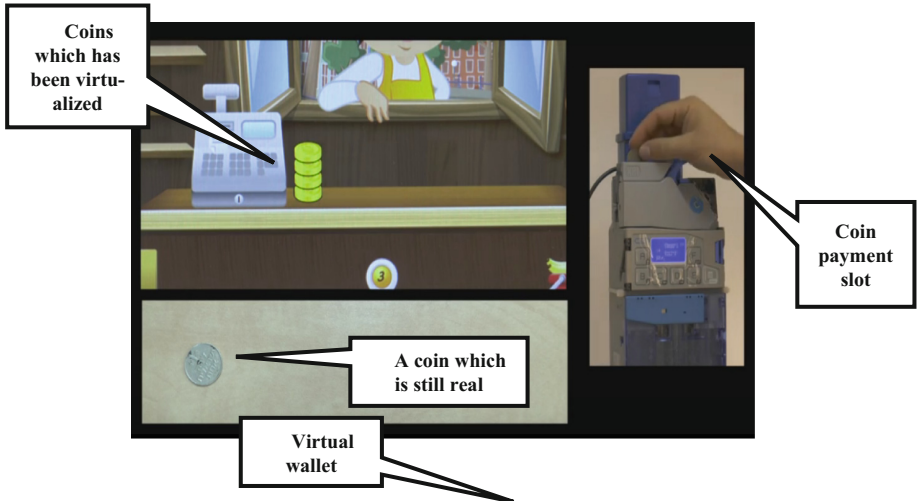


Fig. 3. The experiment system – tangible interface version



Fig. 4. The experiment system – virtual version

In this game the decision the player makes is how many products to buy from the supplier. There were seven events of supplier arrival during the game. Two parameters are considered in this paper: how many items the player decided to buy and how much time it took him to get the decision. The measurement of the bought quantity is made in percentage from the maximum available quantity (it varies from product to product and from one round of the game to another. 100% means the player bought the maximum quantity possible).

4 Results

4.1 Adults Group

The adult participants who played the game with the tangible interface spent less money and it took them more time to make their decisions, compared to the players in the game without the tangible interface. We can see that after a stabilization period, the virtual version’s participants bought about 60% of the maximum quantity and the augmented reality group bought about 45% (Fig. 5).

The decision time period was longer in the game with the tangible interface, as we can see in Fig. 6: about 5 s in the tangible interface version, compared to about 3.5 s while playing the virtual version.

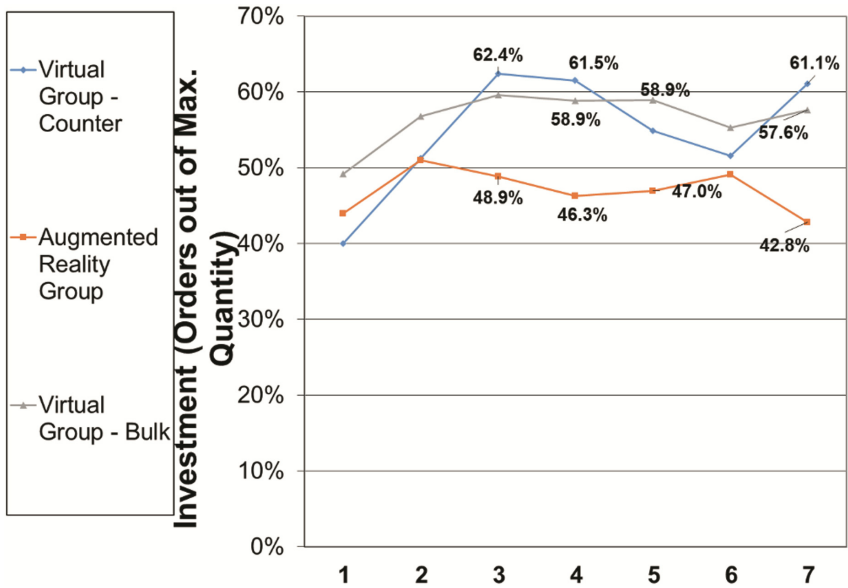


Fig. 5. Results of adults experiment – investment

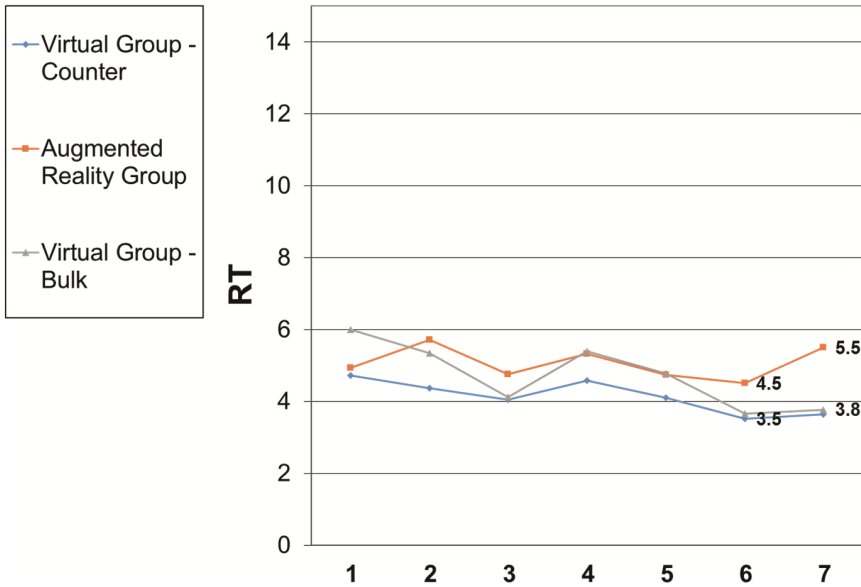


Fig. 6. Results of adults experiment – decision time period

Adults took less risk and their decisions are more considered in the game using the tangible interface.

4.2 Children Group

One sees immediately that the children’s results were totally different from those of the adults: The children that played the game with the tangible interface spent much more money than the children in the reference group. They bought about 70% of the possible quantity when the virtual group’s children bought less than 30% of it (Fig. 7). The decision period was similar in both cases.

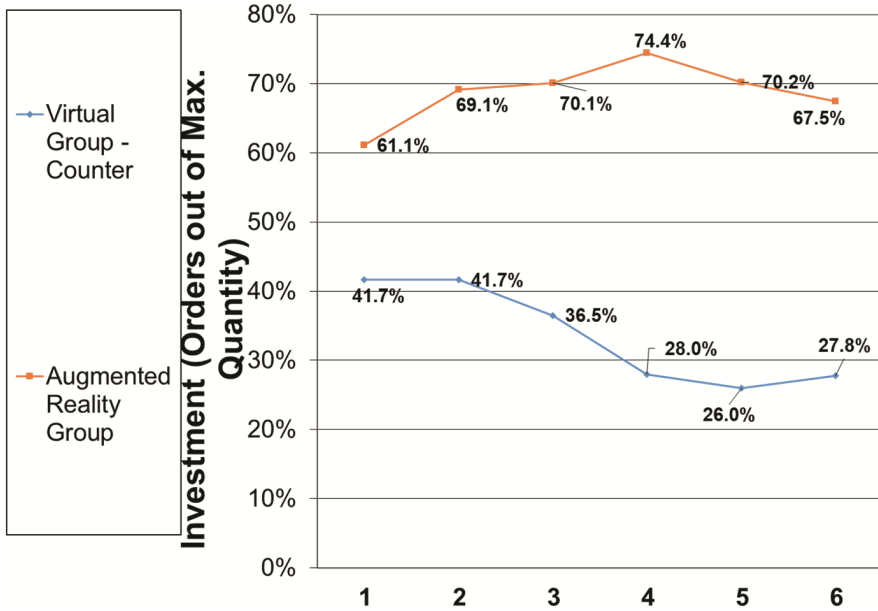


Fig. 7. Results of children experiment – investment

5 Conclusion

The results in the children’s group seem to be the very different from those in the adults group. However, we propose that the psychological reason for the behavior difference caused by using tangible interface within kids and adults is the same.

This experiment verifies well known psychological truths: the usage of the sense of touch increases user attention and involvement in an activity. If it is better to see once than hear a hundred times, then it’s better to touch once than see a hundred times.

We know from cognitive development theory that using the sense of touch is essential for little children, and their ability to deal with abstractions develops later in childhood. The experimental finding verifies this theory: the difference between the children’s groups was much more pronounced. The adults use abstract concepts during their daily life (using a credit card as opposed to cash, for example) and they do it successfully, in general. Children’s ability to use abstraction is limited – from their point of view the virtual game and the game with tangible interface are two different games.

Both adults and children understand better the scenario of the game while using physical coins. This understanding causes the adults to think more before taking decision and to take less risky decisions. The same understanding causes kids to understand the process of buying and to enjoy it. That's why the children bought more while using physical coins. This understanding of the game causes both to adults and to children to do what they want to do: to manage the finances for adults and to enjoy the game for children.

Children are influenced by tangible interface drastically and tangible interface is much more attractive for children.

The game with tangible interface is perceived totally in different way.

6 Possible Implementation

This paper introduces the concept of the tangible interface and examines deeply the usage of tangible interface in children's game. But the tangible interface is not limited to this area only. Other areas that seem to have benefit while using tangible interfaces are gambling platforms and real-task training platforms.

References

1. Milgram, P., Takemura, H., Utsumi, A., Kishino, F.: Augmented reality: a class of displays on the reality-virtuality continuum. ATR Communication Systems Research Laboratoriesf 2-2 Hikaridai, Seika-cho, Soraku-gun Kyoto 619-02 Japan
2. Benford, S., Greenhalgh, C., Reynard, G., Brown, C., Koleva, B.: Understanding and constructing shared spaces with mixed-reality boundaries. *ACM Trans. Comput.-Hum. Interact.* **5**, 185–223 (1998)
3. Van Krevelen, D.W.F., Poelman, R.: A survey of augmented reality technologies, applications and limitations. *Int. J. Virtual Real.* **9**, 1–20 (2010)



Population Stereotypes for Color Associations

Yuting Sun^{1,2} and Kim-Phuong L. Vu²(✉)

¹ Social Security Administration, Baltimore, USA
Yuting.Sun@ssa.gov

² California State University, Long Beach, USA
Kim.Vu@csulb.edu

Abstract. Strong associations of meaning to colors allow designers to use a color not only to capture attention (e.g., use of a bright color to highlight information), but to code meaning (e.g., use of the color red to convey the meaning of danger). The association of meaning to color is learned through experience and, as a result, may differ across cultures. The current study examined population stereotypes for colors by having participants from three different countries (USA, India, and UK) indicate what meaning they associate with 6 common colors. The results showed that the top 1–2 color associations were generally consistent across countries and gender, but the degree of association was different among them. The difference in the degree of color associations by participants from different countries should be taken into account when making design decisions, especially for products that are intended for an international market.

Keywords: Population stereotypes · Color association · Interpretation of colors
Display-control configurations

1 Introduction

The strong association between a display element and its intended meaning can be considered a population stereotype [1]. Population stereotypes are important to design because they can be used to predict how users would interact with a product or system. Early research on population stereotypes focused on participants' natural response tendencies for display-control configurations, such as turning a knob clockwise to increase the value of a display or moving the location of a switch to the up position to turn on a light in a room [2, 3]. Additionally, there was also interest in what meanings individuals associate with certain colors [4]. Bergum and Bergum's [2] 1981 study found that more than 96% of American participants surveyed showed strong associations for the meaning of "go" with the color green, "stop" with the color red, and "cold" with the color of blue. In contrast, the colors of orange and purple were not strongly associated with any meaning. However, color associations found with American participants may not be to the same as those obtained with participants from other countries. Courtney's [4] 1986 study found that less than 65% of the Chinese respondents surveyed associated specific meanings to colors. In fact, only 44.7% of his participants associated "go" with the color green, 48.5% associated "stop" with the color red, and 5.9% associated "cold" with the color of blue. Toriizuka et al. [6] found

that Japanese participants related the colors of red and blue to opposing meanings such as dangerous (reddish hues) vs. safe (blueish hues). The difference observed between cultures in the type and degree of color-to-meaning associations has important implications for the use of color coding in design.

Even within an ethnic culture, the degree of association of a meaning with a color can vary depending on the experiences of participants. In 2001, Chan and Courtney [5] conducted a study to examine color associations for Chinese participants from Hong Kong. In comparison to the Chinese participants from the Yunnan Province of China surveyed in Courtney's previous [4] study, Chan and Courtney found that "go" was more strongly with the color green (62.6%), "stop" with the color red (66.4%), and "cold" with the color of blue (22.5%). Although the level of association between the colors and meaning were higher for participants from Hong Kong compared to Yunnan, the level of association of meaning to color for the Chinese participants were not as strong as those obtained by American participants [2]. The differences in degree of associations found between these three groups of participants points to the need to survey participants from different cultures and locations to determine whether certain color associations apply to a targeted user group.

The present study surveys over 300 participants from three different countries (USA, UK, and India) to capture population stereotypes for common colors. The data presented in this paper is a subset of the larger study that examined population stereotypes for verbal and pictorial displays, and only focuses on population stereotypes for colors.

2 Method

Participants. All participants were adults, 18 years of age or older, and recruited through Amazon Mechanical Turk (Mturk). Participants had to meet the requirements of having an approval rating of 95% or above in Mturk and being a resident of USA, UK, or India. Over 300 responses were collected, with 126 respondents (64 female, 62 male) from the USA, 127 (24 female, 103 male) from India, and 65 (34 female, 30 male, 1 declined to state) from the UK.

Procedure. The online survey was created using Qualtrics and administered to participants through Amazon Mechanical Turk (Mturk). Mturk workers who met the study's qualifications were able to find the survey on a list of HITs (Human Intelligence Tasks) that were available to them as potential participants. These workers became actual participants once they accepted the task and clicked on the link to take our survey. Data collection started once participants accessed the survey through Qualtrics. The first page of the survey was a consent form. Participants acknowledged that they were voluntarily participating in the survey by clicking on a link, which started the survey questions.

As noted earlier, the data presented in this paper is part of a larger study. Participants were presented with 96 questions in randomized order prior to collecting demographic information. All questions had to be answered to complete the survey, but a "decline to answer" option was provided for all questions so that participants could

refuse to answer any question and still participate in the study. For the color association questions, participants were asked to answer the questions by selecting the meaning they associated with a specific color from a list of pre-defined responses. The six colors (red, green, blue, yellow, orange, and purple) used in the survey were the ones that were used in Bergum and Bergum’s [2] 1981 study. The response alternatives provided to participants were also generated from the Bergum and Bergum [2] study; however, we added an option of explicitly indicating that “I don’t associate the color with any of the [meanings] above” (see Fig. 1). Participants were able to select multiple responses to the questions.

Select all options below that you associate with Red.

| | |
|-----------|---|
| Safe | Off |
| Cold | Far |
| Go | Danger |
| Radiation | Caution |
| On | Stop |
| Near | I don't associate the color with any of the above |
| Hot | Decline to answer |

Fig. 1. Example survey question for the color red. Participants were asked to select all response options that they associated with the color red. (Color figure online)

Once the participants completed all the survey questions, they were asked to provide demographic information. The very last question of the survey provided the participants with a verification code to claim their payment of \$0.75 through Mturk for completing the survey.

3 Results and Discussion

Responses to each of the six colors were summarized separately by country and gender. The percentage of responses for each group was reported due to unequal number of participants obtained from different countries and a much larger number of male participants compared to female participants from India. Bergum and Bergum [2] indicated that, for applied research, a response rate of 85% or higher could serve as a criterion for a population stereotype. However, they also noted that stereotypes with lower levels of association (i.e., 66%) could still be useful. Courtney [5] noted that none of the associations in his study reached the 85% level, but some that reached the 60% association rate reached statistical significance. For this paper, we use an intermediate criterion of 75%. However, we leave it to the reader to determine what level of agreement is appropriate for their design purposes.

Red. As shown in Table 1, the top three associations for the color red were the meanings of “Danger”, “Stop”, and “Hot” for participants in all three countries. In terms of reaching the 75% criterion (see Fig. 2), Danger was associated with red for all participants in the present study. Stop was associated with red for all participants, except males from India. Hot was associated with red above the 75% level for female participants from the USA, and both male and female participants from the UK. In other words, UK participants showed the strongest associations for the meanings of “Danger”, “Stop”, and “Hot” with red, and participants from India showed the weakest associations of these concepts with red. In comparison to the prior studies of Bergum and Bergum [2], Courtney [4], and Chan and Courtney [5], the finding of the color red being strongly associated with “Stop” and “Danger” was robust.

Table 1. Top 4 responses for associations with the color red obtained in prior studies, and in the present study from participants in the USA, UK, and India.

| Red | Courtney (1986) Yunnan Chinese | Chan & Courtney (2001) Hong Kong Chinese | Bergum & Bergum (1981) USA | Present Study (2018) USA | Present Study (2018) UK | Present Study (2018) India |
|-----|--------------------------------|--|----------------------------|--------------------------|-------------------------|----------------------------|
| 1 | Danger (64.7%) | Stop (66.4%) | Stop (100%) | Stop (90.4%) | Danger (92.2%) | Danger (82.6) |
| 2 | Stop (48.5%) | Danger (63%) | Hot (94.5%) | Danger (80%) | Hot (90.9%) | Stop (74%) |
| 3 | Caution (37.7%) | Caution (40.2%) | Danger (89.8%) | Hot (73.5%) | Stop (90.8%) | Hot (49.3%) |
| 4 | Hot (31.1%) | Radiation Hazard (27.3%) | Radiation (59.1%) | Off (31.2%) | Caution (58%) | Caution (42.3%) |

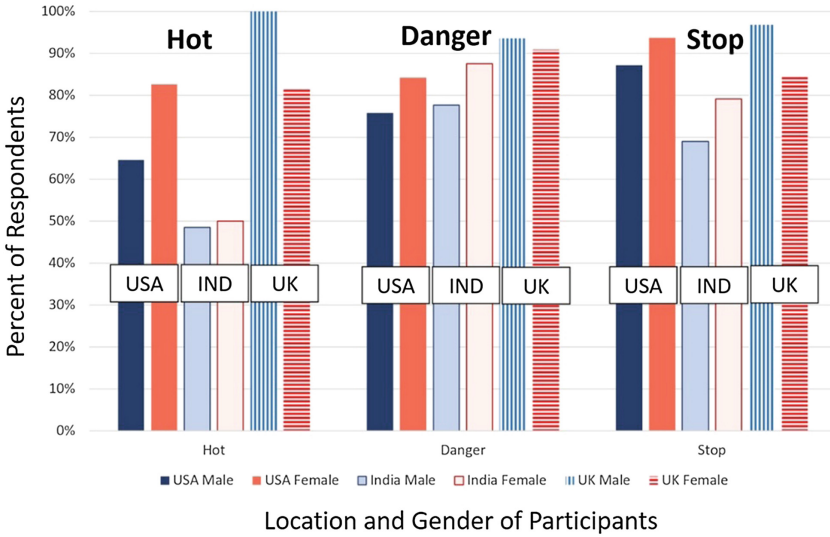


Fig. 2. Top meanings associated with the color red obtained in the present study by location (USA, India, and UK) and gender (male vs. female). (Color figure online)

Table 2. Top 4 responses for associations with the color orange obtained in prior studies, and in the present study from participants in the USA, UK, and India.

| Orange | Courtney (1986) Yunnan Chinese | Chan & Courtney (2001) Hong Kong Chinese | Bergum & Bergum (1981) USA | Present Study (2018) USA | Present Study (2018) UK | Present Study (2018) India |
|--------|--------------------------------|--|----------------------------|--------------------------|-------------------------|----------------------------|
| 1 | Hot (29.8%) | Hot (28.2%) | Near (19.7%) | Caution (58.4%) | Caution (65.8%) | Caution (31.6%) |
| 2 | On (12.3%) | Caution (13.2%) | Far (15%) | Danger (30.4%) | Hot (26.6%) | Hot (20.3%) |
| 3 | Go (8.5%) | Soft (12.3%) | Radiation (13.4%) | Radiation (28%) | Danger (23.5%) | Radiation (18%) |
| 4 | Stop (7.1%) | Potential Hazard (9.5%) | Caution (7.1%) | Hot (16%) | Radiation (17.3%) | Danger (15.1%) |

Orange. As shown in Table 2, the top association for the color orange was the meaning of “Caution” for participants in all three countries of the present study. However, “Caution”, did not reach the 75% criterion for any group of participants. USA and UK participants showed the strongest association of the color orange with “Caution” (58%–71%). The second highest association, “Danger,” yielded less than 31% of responses. The finding of caution being somewhat associated with orange is different from that of Bergum and Bergum [2], Courtney [4], and Chan and Courtney [5] who all found little or no association of “Caution” with the color of orange.

Yellow. As shown in Table 3, “Caution” was the top association for the color yellow for participants from all three countries examined in the present study. For USA participants, the response level exceeded the 75% criterion. The level of association was lower for participants from the UK (52–68%) and India (38–42%). Radiation was the second highest association for participants from India and UK and the third highest association for participants from the USA. However, the level of association was not strong (<40%), see Fig. 3. The finding of yellow being somewhat associated with “Caution” was consistent with prior research studies [2, 4, 5], where “Caution” was also found to be the top or second highest meaning to be associated with the color yellow.

Table 3. Top 4 responses for associations with the color yellow obtained in prior studies, and in the present study from participants in the USA, UK, and India.

| Yellow | Courtney (1986) Yunnan Chinese | Chan & Courtney (2001) Hong Kong Chinese | Bergum & Bergum (1981) USA | Present Study (2018) USA | Present Study (2018) UK | Present Study (2018) India |
|--------|--------------------------------|--|----------------------------|--------------------------|-------------------------|----------------------------|
| 1 | Caution (44.8%) | Radiation Hazard (24.2%) | Caution (81.1%) | Caution (84.8%) | Caution (59.6%) | Caution (39.8%) |
| 2 | Stop (11.6%) | Caution (23.6%) | Near (38.6%) | Danger (24.8%) | Radiation (34.2%) | Radiation (23.7%) |
| 3 | Go (9.7%) | Potential Hazard (17.3%) | Safe (16.5%) | Radiation (20%) | Safe (9.3%) | Near (15.5%) |
| 4 | On (9.6%) | Hot (14.9%) | Radiation (15.7%) | Safe (8.8%) | Hot (8%) | Cold (15.1%) |

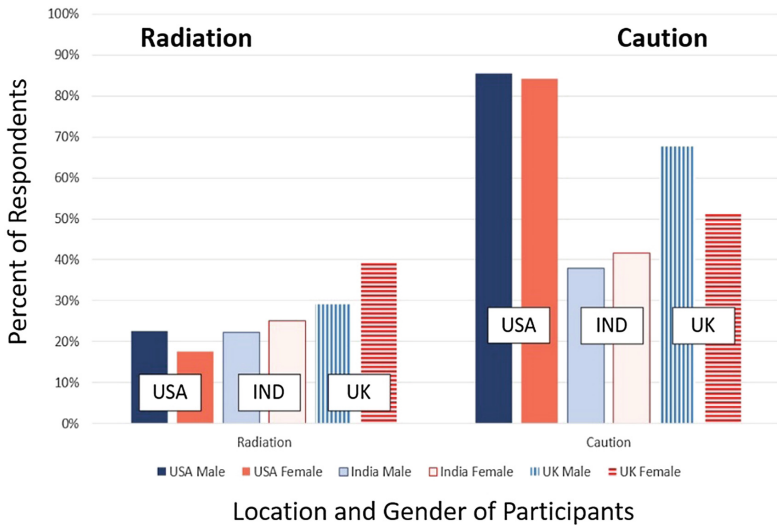


Fig. 3. Top meanings associated with the color yellow obtained in the present study by location (USA, India, and UK) and gender (male vs. female). (Color figure online)

Green. As shown in Table 4, the top three associations for the color green were the meanings of “Go”, “Safe”, and “On” for participants in all three countries examined in the present study and replicates prior research [2, 4, 5]. In terms of reaching the 75% criterion, “Go” was associated with green for both male and female participants from the USA and UK, see Fig. 4. Only “Safe” was highly associated for both male and female participants from the UK. For UK males, “On” was associated strongly with green at 77.4%.

Table 4. Top 4 responses for associations with the color green obtained in prior studies, and in the present study from participants in the USA, UK, and India.

| Green | Courtney (1986) Yunnan Chinese | Chan & Courtney (2001) Hong Kong Chinese | Bergum & Bergum (1981) USA | Present Study (2018) USA | Present Study (2018) UK | Present Study (2018) India |
|-------|--------------------------------|--|----------------------------|--------------------------|-------------------------|----------------------------|
| 1 | Safe (62.2%) | Go (62.6%) | Go (99.2%) | Go (92.8%) | Go (86.1%) | Safe (65.2%) |
| 2 | Go (44.7%) | Safe (38.2%) | Safe (61.4%) | Safe (61.6%) | Safe (84.7%) | Go (63.3%) |
| 3 | On (22.3%) | On (23.8%) | On (37.8%) | On (57.6%) | On (66%) | On (45.9%) |
| 4 | Off (6.3%) | Normal (11.9%) | Off (15%) | Near/ Radiation (4%) | Radiation (11%) | Near (21.2%) |

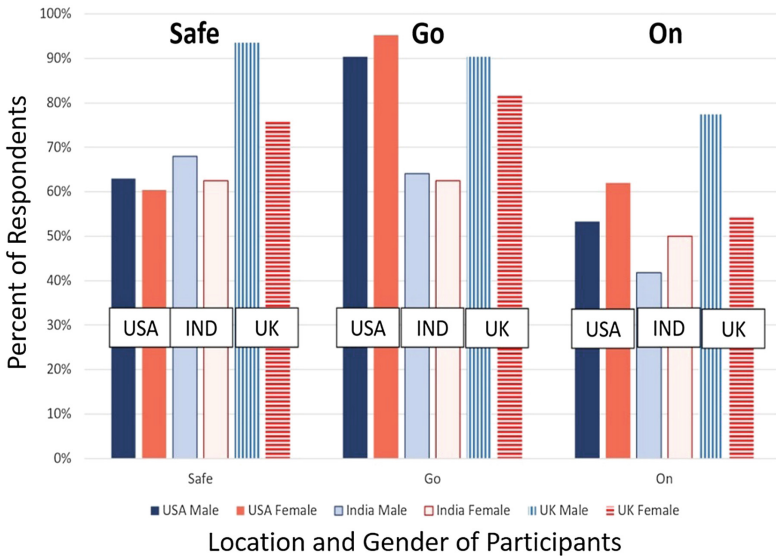


Fig. 4. Top meanings associated with the color green obtained in the present study by location (USA, India, and UK) and gender (male vs. female). (Color figure online)

Blue. As shown in Table 5, the top two associations for blue were the meanings of “Cold” and “Safe” by participants from all three countries in the current study. “Cold” was also the top association with Blue found by Bergum and Bergum [2] and Chan and Courtney [5]. “Safe” was the second highest association with blue found by Courtney [4]. In terms of reaching the 75% criterion, “Cold” was associated with blue for both male and female participants from the USA and UK. “Safe” did not come close to the 75% criterion with association rates of less than 41%, see Fig. 5.

Table 5. Top 4 responses for associations with the color blue obtained in prior studies, and in the present study from participants in the USA, UK, and India.

| Blue | Courtney (1986) Yunnan Chinese | Chan & Courtney (2001) Hong Kong Chinese | Bergum & Bergum (1981) USA | Present Study (2018) USA | Present Study (2018) UK | Present Study (2018) India |
|------|--------------------------------|--|----------------------------|--------------------------|-------------------------|----------------------------|
| 1 | Go (19.9%) | Cold (22.5%) | Cold (96.1%) | Cold (82.3%) | Cold (90.5%) | Cold (40.7%) |
| 2 | Safe (18.4%) | Normal (15.7%) | Off (31.5%) | Safe (20%) | Safe (27%) | Safe (25.6%) |
| 3 | On (14.4%) | Strong (12.8%) | Far (30.7%) | On (13.6%) | Far (18.7%) | Radiation (19.8%) |
| 4 | Off (8.5%) | On (12.7%) | Safe (18.1%) | Near/Go/Radiation (4%) | On (12.6%) | On (18.7%) |

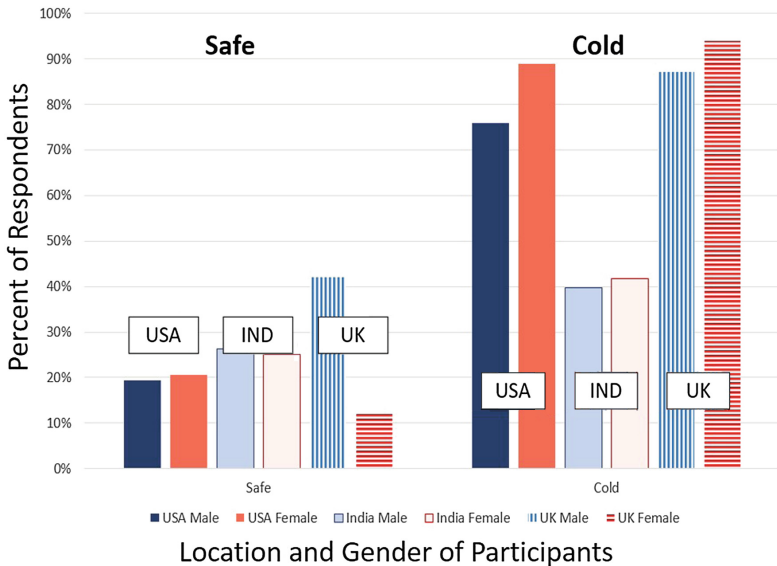


Fig. 5. Top meanings associated with the color blue obtained in the present study location (USA, India, and UK) and gender (male vs. female). (Color figure online)

Purple. There was no stereotypic response for the color purple among participants from all three countries surveyed in this study. Participants from the USA, India and UK were consistent in indicating that they do not associate the color purple with any of the meanings provided. This finding is consistent with prior research [2, 4, 5] in Table 6.

Table 6. Top 4 responses for associations with the color purple obtained in prior studies, and in the present study from participants in the USA, UK, and India.

| Purple | Courtney (1986) Yunnan Chinese | Chan & Courtney (2001) Hong Kong Chinese | Bergum & Bergum (1981) USA | Present Study (2018) USA | Present Study (2018) UK | Present Study (2018) India |
|--------|--------------------------------|--|----------------------------|--------------------------|-------------------------|----------------------------|
| 1 | Hot (17.2%) | Radiation Hazard (22.7%) | Far (34.6%) | None (79.2%) | None (75%) | None (54.4%) |
| 2 | Cold (12%) | Potential Hazard (15.6%) | Off (12.6%) | Cold (11.2%) | Far (10.9%) | Radiation (20.8%) |
| 3 | On (10.3%) | Soft (14.2%) | Radiation (7.9%) | Radiation (8%) | Cold (9.4%) | Cold (14.4%) |
| 4 | Go (9%) | Strong (10.6%) | Near (7.1%) | Safe/Far (3.2%) | Radiation (7.9%) | Far/Caution (8.1%) |

4 Conclusions

Using the 75% criterion, we found that only the color red yielded strong association with the meaning of “Danger” across participants from all three countries surveyed for the present study. The color association of red with “Stop” was the second strongest association. Both participants from the USA and UK indicated strong associations, while participants from India almost met the 75% criterion. In general, participants surveyed from the USA and UK showed stronger associations of meaning with color, compared to participants from India. For example, both USA and UK participants strongly associated the color green with “Go” and blue with “Cold”. Participants from the UK also strongly associated the color red with “Stop” and green with “Safe”, while USA participants strongly associated yellow with “Caution”. Participants from India showed only two associations of color to meaning above the 70% level (red with “Danger” and “Stop”) and 2 at the response rate of 60% or higher (green with “Go” and “Safe”). The finding of lower rates of color-to-meaning associations for participants from India in comparison with US participants is consistent to what Courtney [4] and Chan and Courtney [5] found for Chinese participants in comparison to US participants.

The robust association of red with “Danger” is one that designers can use with confidence when using color to convey warnings for products. Care should be taken when using other colors to convey meaning as the degree of association may not be strong across cultures. Our findings indicate that the color of purple should be avoided

when conveying meaning for products aimed at general users, as it was not strongly associated to any meaning examined in the present study.

The present study also found some results that were consistent with the prior studies of Bergum and Burgum [2], Courtney [4], and Chan and Courtney [5], indicating the robustness of color-to-meaning associations over time. However, there were disparities that could reflect differences in experiences between cultures or changes over time. If colors are used in design to convey specific meanings, it is recommended that the designers verify the degree of the color-to-meaning association with participants from the targeted user group(s). In addition, designers need to pay extra care when using colors to convey meaning for products that are designed for international markets.

Acknowledgments. We thank Allen Chen, Matthew Nare, and Sabrina Moran from the Center of Usability in Design and Accessibility (CUDA) in their assistance to this project.

References

1. Proctor, R.W., Vu, K.-P.L.: Biological, ecological, and cultural contributions to display-control compatibility. *Am. J. Psychol.* **123**, 425–435 (2010)
2. Bergum, B.O., Bergum, J.E.: Population stereotypes: an attempt to measure and define. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* vol. 25, No. 1, pp. 662–665. Sage Publications (1981)
3. Smith, S.L.: Exploring compatibility with words and pictures. *Hum. Factors* **23**, 305–315 (1981)
4. Courtney, A.J.: Chinese population stereotypes: color associations. *Hum. Factors* **28**, 97–99 (1986)
5. Chan, A.H.S., Courtney, A.J.: Color associations for Hong Kong Chinese. *Int. J. Ind. Ergon.* **28**, 165–170 (2001)
6. Toriizuka, T., Ikeda, M., Löffler, D., Hurtienne, J.: Conveying opposite system states through Color–Japanese population stereotypes of Hue–Concept associations. In: *Proceedings 19th Triennial Congress of the IEA*, vol. 9, p. 14 (2015)



Presentation of Personal Health Information for Consumers: An Experimental Comparison of Four Visualization Formats

Da Tao¹, Juan Yuan¹, Xingda Qu¹, Tiejian Wang¹, and Xingyu Chen^{2(✉)}

¹ Institute of Human Factors and Ergonomics, Shenzhen University, Shenzhen, China
{taoda, quxd}@szu.edu.cn, 2515809640@qq.com,
wangtiejian188@sina.com

² Department of Marketing, Shenzhen University, Shenzhen, China
celine@szu.edu.cn

Abstract. While the development of consumer-oriented health information technologies (CHITs) has led to increased availability and accessibility of personal health information, consumers may encounter difficulty in comprehending the information, partly due to inappropriate information presentation. This study was conducted to compare four visualization formats of personal health information in consumers' use and comprehension of the information. A within-subjects design was employed, with visualization format serving as independent variable, and sets of user performance, perception, eye movement and preference measures serving as dependent variables. Twenty-four participants were recruited in this study. The results indicated that there was no significant main effect of visualization format on task completion time and accuracy rate, while visualization format yielded a significant effect on perceived health risk, perceived ease of understanding, perceived usefulness, perceived confidence of comprehension, and satisfaction. Participants' visual attention, indicated by eye movement measures, was significantly affected by areas of interest, but not by visualization format. Most participants preferred personalized enhanced format. Our study demonstrates that visualization formats could affect how personal health information are comprehended and perceived. The results may help to improve the design of more usable and effective health information presentation.

Keywords: Visualization format · Health information · Comprehension Presentation

1 Introduction

The healthcare domain is facing great challenge due to increasing demands on healthcare services from people with chronic diseases and suboptimal health status, and the ageing population. In China, 300 million people are suffering from various chronic diseases [1], 1030 million people are experiencing suboptimal health status [2] and 220 million people are aged 60 years or above [3]. Research has shown that one of effective approaches to meet consumers' healthcare demands is continuous self-monitoring of health indicators

(e.g., heart rate, blood pressure and blood glucose), which can be facilitated by consumer-oriented health information technologies (CHITs) [4–6]. CHITs refer to consumer-centered electronic tools, technologies, applications, or systems that are interacted with directly by health consumers (i.e., individuals who seek or receive health care services) to provide them with data, information, recommendations, or services for promotion of health and health care [4, 6]. CHITs are convenient tools to track, record and manage consumers' personal health information (e.g., blood pressure), and can easily present the information for a wide range of consumers [7–9].

While the development of CHITs has led to increased availability and accessibility of personal health information, consumers may encounter difficulty in comprehending and thus correctly responding to the information, partly due to inappropriate information presentation [10–12]. This is a significant concern in health care, as inappropriate presentation of health information may lead to confusion, frustration and disruption in consumers' healthcare process [13, 14] and even to adverse consequences, such as medication error and inappropriate healthcare decision-making [15, 16]. In fact, there is much evidence that consumers find it difficult to understand quantitative health information [17–19]. This is especially the case for people with low numeracy and literacy skills [20, 21]. While many consumers are in urgent need of understanding their health status, we know little about optimal presentation of personal health information for them.

The way health information presented can have significant influence on what the information is processed, a phenomenon known as the representational effect [22]. It has been increasingly recognized that the use of visualization may be an effective way to present quantitative health information, and is likely to improve interpretation and comprehension of the information [11, 23]. For example, Torsvik et al. found that visualization formats, such as sparklines and relative multigraphs, seem to be favorable techniques for presenting complex long-term clinical test results, while tables seem better for simpler test results [11]. However, until relatively recently, there has been little research to inform which kinds of visualization formats are optimal to support consumers' use and comprehension of personal health information. There is also a lack of research to describe how consumers perceive different visualization formats for their personal health information (e.g., whether a particular type of visualization format is perceived helpful or not in their healthcare).

The purpose of this study was to evaluate four visualization formats in consumers' use and comprehension of personal health information. The visualization formats were applied to two types of personal health information, i.e., blood pressure and blood glucose, which are main indicators that are usually monitored by chronically ill patients (especially those with hypertension and/or diabetes) [5, 7].

2 Methods

2.1 Experimental Design

A within-subjects design was employed, with visualization format (Four types: basic format, color format, color/text format and personalized format) serving as an independent variable, and sets of user performance (i.e., task completion time and accuracy

rate), eye movement (i.e., time to first fixation and total fixation duration), perception (i.e., perceived health risk, perceived ease of understanding, perceived usefulness, perceived confidence and satisfaction) and preference measures serving as dependent variables. Task completion time referred to the total time a participant spent to answer question in a specific task. Accuracy rate was calculated as the proportion of answers that were correctly answered for one type of visualization format. Eye movement measures were assessed to examine visual attention during task performance and were recorded using a Tobii X-120 eye tracker (Tobii Technology, Sweden). User preference was assessed by asking participants to choose their most preferred visualization format.

2.2 Participants

Twenty-four students (12 males and 12 females; mean age 22.1 years (SD 2.4)) participated in this study. They all had self-reported normal color vision and basic numeric knowledge and literacy. A minimal sample size of 17 was required to detect a medium effect size of 0.3 between visualization formats when statistical power and level of significance were selected at 80% and 5%, respectively. The study protocol was approved by the Institutional Review Board of Shenzhen University. Informed consent was obtained from each of the participants.

2.3 Materials and Tasks

All the four visualization formats were applied to results for two types of self-monitoring tests. self-monitoring of blood pressure presented results for diastolic and systolic blood pressure, while self-monitoring of blood glucose presented results for fasting blood glucose and two hours postprandial blood glucose.

All the four visualization formats were created based on horizontal bar graphs, which are commonly applied for displaying individual test results [10, 13]. Information presented in the visualization formats included test name, exact test value, unit of measurement, and cut-off points for normal range. Reference information of normal range for the test results was provided and put at the bottom of the graphs. The four visualization formats were described as follows. Basic format used non-color bar only. Color enhanced format applied color on the basic format, with green and red indicating normal and abnormal range, respectively; but the color meaning was not explained. Text/color enhanced format, based on the design of color enhanced format, provided explicit text explanation for the color to indicate whether the test result was normal or not. Personalized enhanced format, based on the design of color/text enhanced format, provided additional personalized information that was assumed to be an average value of the test results from population with the same sex and age as the participants (See Fig. 1 for an example). Four areas of interest (AOIs) were drawn for each graph to examine participants' visual attention. The first three AOIs covered area that presented different information from basic format, while the fourth one covered the area of reference information of normal range.

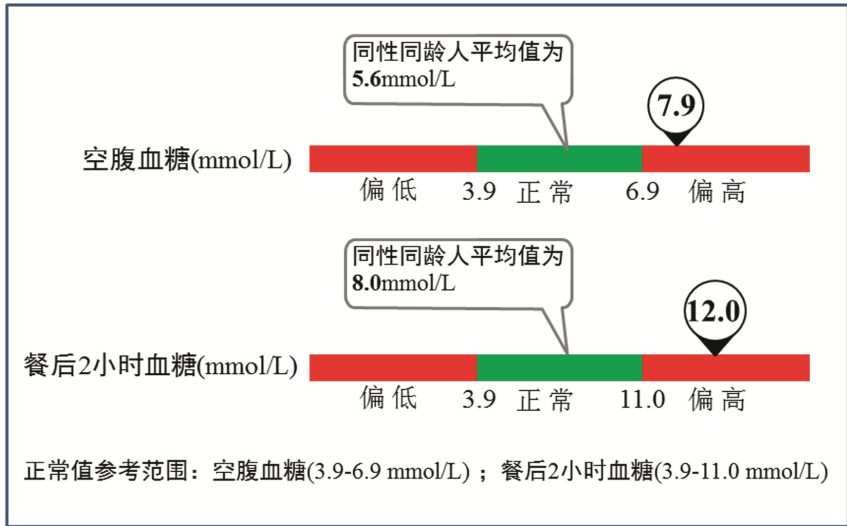


Fig. 1. Example of personalized enhanced visualization format for test results of self-monitoring of blood glucose. (Color figure online)

The experimental test included two types of tasks. Information search tasks asked participants to answer the exact test value present on the visualization format, while judgement tasks required participants to indicate whether the test value was normal or not.

2.4 Procedures

Task scenarios were performed on a DELL computer (Screen size: 23 inches; resolution: 1024 × 768). The eye tracker was equipped at the lower edge of the computer screen. Before the experiment, participants provided informed consent and were given detailed information of test procedures. The participants were then instructed to sit at a fixed distance from the computer screen, and to follow standard eye tracker calibration procedures. Following several practice tasks to familiarize themselves with the test, participants were asked to initiate the main experimental tasks. Participants were asked to respond as quickly and accurately as possible. Combinations of visualization format, and type of test were randomized in a full factorial design. After the experimental tasks, participants were required to complete a paper-based questionnaire to elicit their response to perception measures and preference. The whole experiment took approximately 40 min.

2.5 Data Analysis

Repeated measures analyses of variance (ANOVAs) were used to analyze the effects of visualization format on user performance, eye movement, and perception measures. Past

hoc analyses were performed with Bonferroni adjustment where necessary. Chi-square test was performed to examine the difference in user preference. Level of significance was set at $\alpha = 0.05$. Statistical analyses were performed using SPSS 22.

3 Results

3.1 Performance Measures

Table 1 presents ANOVA analysis results for task completion time and accuracy rate. There was no significant main effect of visualization format on task completion time ($F(3, 60) = 2.09, p = 0.111$) and accuracy rate ($F(3, 60) = 1.31, p = 0.280$) in information search tasks. Similarly, visualization format had no effect on task completion time ($F(1.947, 38.938) = 1.48, p = 0.240$) and accuracy rate ($F(1.679, 33.579) = 1.16, p = 0.331$) in judgement tasks.

Table 1. Effects of visualization format on task completion time and accuracy rate.

| Visualization format | Task completion time (s) | | | | Accuracy rate (%) | | | |
|--------------------------------|--------------------------|-----|---------|---------|----------------------|-----|---------|---------|
| | Descriptive analysis | | ANOVA | | Descriptive analysis | | ANOVA | |
| | Mean | SD | F value | p value | Mean | SD | F value | p value |
| Information search task | | | | | | | | |
| Basic | 14.5 | 5.0 | 2.09 | 0.111 | 88.7 | 0.3 | 1.31 | 0.280 |
| Color | 14.6 | 5.3 | | | 89.6 | 0.3 | | |
| Color/text | 12.3 | 5.1 | | | 85.5 | 0.3 | | |
| Personalized | 16.0 | 6.6 | | | 82.1 | 0.4 | | |
| Judgement task | | | | | | | | |
| Basic | 3.0 | 0.9 | 1.48 | 0.240 | 90.0 | 0.2 | 1.16 | 0.331 |
| Color | 3.1 | 1.3 | | | 93.0 | 0.2 | | |
| Color/text | 2.6 | 0.6 | | | 88.8 | 0.2 | | |
| Personalized | 3.2 | 1.2 | | | 87.1 | 0.2 | | |

3.2 Perception Measures

Visualization format yielded significant effects on perceived health risk ($F(3, 60) = 2.97, p = 0.040$), perceived ease of understanding ($F(3, 60) = 19.84, p < 0.001$), perceived usefulness ($F(3, 60) = 14.72, p < 0.001$), perceived confidence ($F(2.275, 45.497) = 15.21, p < 0.001$), and satisfaction ($F(2.191, 43.815) = 47.37, p < 0.001$). Perceived health risk was higher for personalized enhanced format than for basic format. Formats with more information cues resulted in more perceived ease of understanding and perceived usefulness, and higher levels of perceived confidence and satisfaction (Table 2).

Table 2. Effects of visualization formats on perception measures.

| Measures | Visualization format | Descriptive analysis | | ANOVA | |
|---------------------------------|----------------------|----------------------|-----|---------|---------|
| | | Mean | SD | F value | p value |
| Perceived health risk | Basic | 3.3 | 1.0 | 2.97 | 0.040 |
| | Color | 3.4 | 1.1 | | |
| | Color/text | 3.5 | 1.0 | | |
| | Personalized | 3.8 | 1.2 | | |
| Perceived ease of understanding | Basic | 3.3 | 1.8 | 19.84 | <0.001 |
| | Color | 4.9 | 1.5 | | |
| | Color/text | 5.6 | 0.8 | | |
| | Personalized | 5.6 | 1.3 | | |
| Perceived usefulness | Basic | 4.5 | 1.6 | 14.72 | <0.001 |
| | Color | 5.5 | 1.6 | | |
| | Color/text | 6.0 | 1.2 | | |
| | Personalized | 6.4 | 1.1 | | |
| Perceived confidence | Basic | 4.2 | 1.8 | 15.21 | <0.001 |
| | Color | 5.4 | 1.6 | | |
| | Color/text | 5.9 | 1.2 | | |
| | Personalized | 6.3 | 0.9 | | |
| Satisfaction | Basic | 3.0 | 1.7 | 47.37 | <0.001 |
| | Color | 5.4 | 1.9 | | |
| | Color/text | 6.2 | 0.8 | | |
| | Personalized | 6.3 | 1.0 | | |

3.3 Eye Movement Measures

Time to first fixation was significantly affected by AOI ($F(3, 21) = 4.87, p = 0.010$), but not by visualization formats ($F(3, 21) = 2.67, p = 0.074$) (Table 3). Both AOI 4 yielded longer time to first fixation than other AOIs. Similarly, total fixation duration was significantly affected by AOI ($F(3, 21) = 15.68, p < 0.001$) but not by visualization formats ($F(3, 21) = 2.11, p = 0.130$). AOI 1 obtained longer total fixation duration than other AOIs (all p 's < 0.05).

3.4 User Preference

Table 4 shows the user preference data on visualization format. Most participants preferred personalized enhanced graph (70.8%, $\chi^2 = 15.75, p < 0.001$).

Table 3. Effects of visualization formats and area of interest on time to first fixation and total fixation duration.

| Visualization format | Time to first fixation (s) | | | | Total fixation duration (s) | | | |
|-----------------------------|----------------------------|-----|---------|---------|-----------------------------|-----|---------|---------|
| | Descriptive analysis | | ANOVA | | Descriptive analysis | | ANOVA | |
| | Mean | SD | F value | p value | Mean | SD | F value | p value |
| <i>Visualization format</i> | | | | | | | | |
| Basic | 2.2 | 0.7 | 2.67 | 0.074 | 0.7 | 0.6 | 2.11 | 0.130 |
| Color | 2.7 | 2.3 | | | 0.7 | 0.7 | | |
| Color/text | 1.7 | 0.5 | | | 0.5 | 0.3 | | |
| Personalized | 2.7 | 1.4 | | | 1.0 | 1.1 | | |
| <i>AOI</i> | | | | | | | | |
| AOI 1 | 1.9 | 0.8 | 4.87 | 0.010 | 1.1 | 0.7 | 15.68 | <0.001 |
| AOI 2 | 2.9 | 1.1 | | | 0.4 | 0.4 | | |
| AOI 3 | 2.6 | 1.6 | | | 0.6 | 0.5 | | |
| AOI 4 | 4.1 | 1.5 | | | 0.7 | 0.5 | | |

AOI, area of interest.

Table 4. Distribution of participant preference by visualization format.

| Visualization format | Percentage |
|----------------------|------------|
| Basic | 0% |
| Color | 8.3% |
| Color/text | 20.8% |
| Personalized | 70.8% |

4 Discussion

CHITs have enabled consumers to get access to their own health records from various self-monitoring tests more frequently. However, poorly designed presentation of test results usually leads to misunderstanding and confusion for consumers, in inefficiency and disruption in their health care process, and in a higher likelihood of committing errors in their medical decision-making. In light of this, the present study evaluate four different visualization formats to explore optimal presentation of personal health information for consumers. This study demonstrates that there are differences between visualization techniques with respect to how personal health information are viewed, possessed and comprehended, and how fast and effectively the comprehension is made.

4.1 Primary Findings

Our study represents a rare attempt to evaluate various visualization formats for personal health information. On one hand, the results show that the presentation of self-monitoring results in different formats had different effects on how consumers evaluated the information. This is congruent with findings from previous studies [10, 11, 23–28].

Consumers considered visualization formats that contained more information cues more useful and easier to understand, and developed more confidence in understanding their self-monitoring results with such formats. In particular, formats that used color/text, or personalized information were favored most by consumers. The findings appear to confirm the effectiveness of color, text and personalized information cues in facilitating consumers' comprehension of self-monitoring results. For example, color format is able to provide consumers immediate and strong impression of whether test values were within normal ranges [11]. Similarly, text and personalized information may work as redundancy check, and thus are likely to support consumers' decision-making in their information comprehension.

One the other hand, we observed only little variation in task performance between basic format and three other formats. This may be due to that differences between the four formats were not sufficient to influence consumers' efficiency and effectiveness in performing healthcare tasks, as they were all designed based on similar graphs, with similar structure and layout. However, this may also imply that the use of varied additional information cues would not cause additional cognitive workload for consumers, though more information needs to be processed.

We found that consumers perceived higher risk for their health status, as more information cues were applied in the visualization formats. This may be that consumers became more cautious and conservative in the evaluation of their health information, and thus consider themselves in a higher risk level, as the visualization formats contained more information. However, it should be noted that there is little consensus regarding which level of health risk is appropriate and should be conveyed to consumers for certain health information. Thus, it remains unclear how information should be visualized to convey appropriate perceived health risk for consumers. Intriguingly,

The present study provided preliminary yet unique evidence on visual attention when consumers view the graphs, which is less investigated in previous studies but particularly important in the visualization of health information. We found that as more information cues were applied in the visualization formats, shorter time to first fixation and longer total fixation time were observed in corresponding AOIs, indicating an attraction effect of information cue. The attraction effect was especially obvious when the color cue was introduced. Moreover, we observed that reference information was less noted. Also, for those who noted the information, it took them longer time to do so. This implies that the current presentation of reference information might need revision, as it was even not noted by consumers. More efforts are required to design innovative ways to present test results and reference information together in a holistic way.

While the importance of user experience measures, such as subjective perceptions and preference, is increasingly recognized in the design of informatics tools, they are largely overlooked in existing literature and information visualization guidelines [29]. This study demonstrated that the majority of participants preferred personalized visualization format. It appears that users favored presentation format that was able to convey better perceptual feelings. User preference is important, as users may largely base their decision of using certain informatics tools on subjective perceptions and preference. Therefore, researchers and practitioners should pay sufficient attention to user preference

in future revision of information visualization guidelines in addition to performance and perception measures.

4.2 Implications

Our findings have important implications for the visualization of personal health information for consumers. Theoretically, our study emphasized the importance of appropriate design of visualization format to improve consumers' performance and comprehension of health information. From a practical perspective, our results are not clear on what is the optimal visualization format for personal health information with respect to how quickly the results could be correctly interpreted. Rather, our study shows advantages and disadvantages of different visualization formats. Providers and designers need to be aware of the differential effects on consumers' comprehension, perceptions, visual attention and preference that may be generated through the use of different visualization formats.

4.3 Limitations

This study has several limitations. First, the generalizability of our findings remains to be established. Our conclusions about the effects of visualization format should be viewed as tentative, as only a limited number of visualization formats were evaluated in our study with a small sample size. Second, while our study was conducted in a controlled laboratory, it did not fully simulate actual use of personal health information. This approach may lead to limited ecological validity of the findings. It is likely that participants might respond differently in a real situation. Finally, we did not address age-related factors, such as health literacy, graph literacy, and cognition ability, which are suggested to affect comprehension [28, 30]. Studies with chronically ill patients, or people with low health literacy and education level may yield different results.

5 Conclusions

It is essential for consumers to accurately comprehend personal health information in their healthcare activities. This study demonstrated that different techniques for visualizing and presenting personal health information influenced on how the information was assessed, perceived and comprehended. More development has to be undertaken to improve the visualization techniques and examine them in practical settings where consumers actually use them in real self-care activities.

References

1. National Health and Family Planning Commission of the PRC: Report on the status of Chinese residents' nutrition and chronic diseases (2015). http://www.nhfpc.gov.cn/jkj/pgzdt/new_list_9.shtml
2. Pan, J.H., Shan, J.J.: Annual Report on Urban Development of China-No. 9. Social Science Academic Press, Beijing (2016)
3. The State Council of the People's Republic of China: National population development plan (2016–2030 year) (2017). http://www.gov.cn/zhengce/content/2017-01/25/content_5163309.htm
4. Tao, D., Wang, T., Wang, T., Liu, S., Qu, X.: Effects of consumer-oriented health information technologies in diabetes management over time: a systematic review and meta-analysis of randomized controlled trials. *J. Am. Med. Inform. Assoc.* **24**(5), 1014–1023 (2017). <https://doi.org/10.1093/jamia/ocx014>
5. Or, C.K.L., Tao, D.: Does the use of consumer health information technology improve outcomes in the patient self-management of diabetes? a meta-analysis and narrative review of randomized controlled trials. *Int. J. Med. Inform.* **83**, 320–329 (2014). <https://doi.org/10.1016/j.ijmedinf.2014.01.009>
6. Tao, D., Shao, F., Liu, S., Wang, T., Qu, X.: Predicting factors of consumer acceptance of health information technologies: a systematic review. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 60(1), pp. 598–602 (2016)
7. Or, C., Tao, D.: A 3-month randomized controlled pilot trial of a patient-centered, computer-based self-monitoring system for the care of type 2 diabetes mellitus and hypertension. *J. Med. Syst.* **40**(4), 81 (2016). <https://doi.org/10.1007/s10916-016-0437-1>
8. Tao, D., Or, C.K.: Effects of self-management health information technology on glycaemic control for patients with diabetes: a meta-analysis of randomized controlled trials. *J. Telemed. Telecare.* **19**, 133–143 (2013). <https://doi.org/10.1177/1357633x13479701>
9. Tao, D., Xie, L.Y., Wang, T.Y., Wang, T.S.: A meta-analysis of the use of electronic reminders for patient adherence to medication in chronic disease care. *J. Telemed. Telecare* **21**(1), 3–13 (2015)
10. Brewer, N.T., Gilkey, M.B., Lillie, S.E., Hesse, B.W., Sheridan, S.L.: Tables or bar graphs? Presenting test results in electronic medical records. *Med. Decis. Making* **32**(4), 545–553 (2012). <https://doi.org/10.1177/0272989x12441395>
11. Torsvik, T., Lillebo, B., Mikkelsen, G.: Presentation of clinical laboratory results: an experimental comparison of four visualization techniques. *J. Am. Med. Inform. Assoc.* **20**(2), 325–331 (2013). <https://doi.org/10.1136/amiajnl-2012-001147>
12. Jimison, H., Gorman, P., Woods, S., Nygren, P., Walker, M., Norris, S., et al.: Barriers and drivers of health information technology use for the elderly, chronically ill, and underserved. Evidence Report/Technology Assessment No. 175 (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024). AHRQ Publication No. 09-E004. Rockville, Agency for Healthcare Research and Quality (2008)
13. Or, C.K.L., Tao, D.: Usability study of a computer-based self-management system for older adults with chronic diseases. *JMIR. Res. Protoc.* **1**, e13 (2012)
14. Tao, D., Or, C. (eds.) A Paper Prototype Usability Study of a Chronic Disease Self-management System for Older Adults. 2012 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 10–13 Dec 2012, Hong Kong (2012)
15. Middleton, B., Bloomrosen, M., Dente, M.A., Hashmat, B., Koppel, R., Overhage, J.M., et al.: Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. *J. Am. Med. Inform. Assoc.* **20**(e1), e2–e8 (2013)

16. Trevena, L.J., Zikmund-Fisher, B.J., Edwards, A., Gaissmaier, W., Galesic, M., Han, P.K., et al.: Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. *BMC Med. Inform. Decis. Mak.* **13**(Suppl 2), S7 (2013). <https://doi.org/10.1186/1472-6947-13-s2-s7>
17. Zikmund-Fisher, B.J., Exe, N.L., Witteman, H.O.: Numeracy and literacy independently predict patients' ability to identify out-of-range test results. *J Med Internet Res.* **16**, e187 (2014)
18. Smith, S.G., Curtis, L.M., O'Connor, R., Federman, A.D., Wolf, M.S.: ABCs or 123 s? The independent contributions of literacy and numeracy skills on health task performance among older adults. *Patient Educ. Couns.* **98**(8), 991–997 (2015)
19. Galesic, M., Garcia-Retamero, R.: Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples. *Arch. Intern. Med.* **170**(5), 462–468 (2010). <https://doi.org/10.1001/archinternmed.2009.481>
20. Serper, M., Patzer, R.E., Curtis, L.M., Smith, S.G., O'Connor, R., Baker, D.W., et al.: Health literacy, cognitive ability, and functional health status among older adults. *Health Serv. Res.* **49**(4), 1249–1267 (2014). <https://doi.org/10.1111/1475-6773.12154>
21. Taha, J., Sharit, J., Czaja, S.J.: The impact of numeracy ability and technology skills on older adults' performance of health management tasks using a patient portal. *J. Appl. Gerontol.* **33**(4), 416–436 (2014)
22. Zhang, J., Norman, D.A.: Representations in distributed cognitive tasks. *Cogn. Sci.* **18**(1), 87–122 (1994)
23. Kurtzman, E.T., Greene, J.: Effective presentation of health care performance information for consumer decision making: A systematic review. *Patient Educ. Couns.* **99**(1), 36–43 (2016). <https://doi.org/10.1016/j.pec.2015.07.030>
24. Hildon, Z., Allwood, D., Black, N.: Impact of format and content of visual display of data on comprehension, choice and preference: a systematic review. *Int. J. Qual. Health Care* **24**(1), 55–64 (2012). <https://doi.org/10.1093/intqhc/mzr072>
25. Timmermans, D.R., Ockhuysen-Vermey, C.F., Henneman, L.: Presenting health risk information in different formats: the effect on participants' cognitive and emotional evaluation and decisions. *Patient Educ. Couns.* **73**(3), 443–447 (2008). <https://doi.org/10.1016/j.pec.2008.07.013>
26. Okan, Y., Stone, E.R., Bruine de Bruin, W.: Designing graphs that promote both risk understanding and behavior change. *Risk Anal.* (2017). <https://doi.org/10.1111/risa.12895>
27. Harris, R., Noble, C., Lowers, V.: Does information form matter when giving tailored risk information to patients in clinical settings? A review of patients' preferences and responses. *Patient Prefer Adherence* **11**, 389–400 (2017). <https://doi.org/10.2147/ppa.s125613>
28. Okan, Y., Galesic, M., Garcia-Retamero, R.: How people with low and high graph literacy process health graphs: evidence from eye-tracking. *J. Behav. Decis. Mak.* **29**(2–3), 271–294 (2016). <https://doi.org/10.1002/bdm.1891>
29. Tao, D., Yuan, J., Liu, S., Qu, X.: Effects of button design characteristics on performance and perceptions of touchscreen use. *Int. J. Ind. Ergon.* **64**, 59–68 (2018). <https://doi.org/10.1016/j.ergon.2017.12.001>
30. Garcia-Retamero, R., Cokely, E.T.: Communicating health risks with visual aids. *Curr. Dir. Psychol. Sci.* **22**(5), 392–399 (2013)



Micro and Macro Predictions: Using SGOMS to Predict Phone App Game Playing and Emergency Operations Centre Responses

Robert West¹(✉), Lawrence Ward², Kate Dudzik¹, Nathan Nagy¹,
and Fraydon Karimi¹

¹ Institute of Cognitive Science, Carleton University, Ottawa, Canada
{robert_west, Kate.dudzik, Nathan.nagy}@carleton.ca,
Fraydon.karimi@Carleton.ca

² Department of Psychology, University of British Columbia,
Vancouver, Canada
lward@psych.ubc.ca

Abstract. In this study, we examine the ability of SGOMS models to predict human behaviour on two different scales, in micro cognitive task performance and in high level problem solving roles to better understand strategy use and training. To do this, two experiments were designed to isolate the role of knowledge structures in task performance. The first experiment involves modelling an application-based game, played on mobile phones. Results were compared to two models: the SGOMS model that matched the knowledge structures the players had learned during training, and a model optimized for speed, resulting in the fastest game play possible using ACT-R. In the second experiment we examined SGOMS predictions in a high level problem space of an Emergency Operations Center (EOC) simulation, with many interruptions and communication demands, comparing professional EOC managers and undergraduate performance. By comparing results between tasks, HCI design can be augmented using predictive modeling to inform the design to produce efficient and effective training programs.

Keywords: HCI · Training · SGOMS · ACT-R · App

1 Introduction

In this paper, we discuss the macro architecture hypothesis (West et al. 2013) and its significance for understanding the role that the cognitive sciences can play in designing systems for use in business, services, and institutions. Specifically, we are concerned with the extent that cognitive modelling can be meaningfully applied to real world systems design. In this study, we explored this by modeling two very different tasks, a simple memory game and a simulated Emergency Operations Center (EOC) response to disasters. By doing this we highlight the differences between high and low level tasks and how they impact modeling.

The idea of the macro architecture hypothesis came out of a debate within the macro cognition research community concerning the value of cognitive psychology

research. The distinction between micro and macro cognition (Cacciabue and Hollnagel 1995; Klein et al. 2003) was created to distinguish complex, real world cognition (macro cognition) from the artificial and simplified scenarios used in Cognitive Psychology experiments (micro cognition). The basic idea was that real world (macro) cognition needs to be studied and understood on its own terms. Since the goal of Cognitive Psychology is to isolate and study fundamental cognitive functions, we naturally assume that the study of macro cognition should be based on the findings of micro cognition. However, some (e.g., Klein et al. 2002) have questioned the value of micro cognition, claiming it does not scale up and is therefore of limited use in macro level system design.

The question underlying this skepticism is whether artificial experiments that do not represent the full complexity of real world cognition tell us anything useful about cognition in the real world. In fact, this general concern is not unique to Macro Cognition. It has been floated by numerous groups concerned with scaling up from lab based experimental results (e.g., Gregson 1988; Kingstone et al. 2003; Turvey and Carello 2012; van Gelder and Port 1995). The macro architecture hypothesis is a response to this criticism that seeks to maintain the use of traditional cognitive psychology, but also addresses concerns about real world complexity.

In his famous paper, *You can't play 20 questions with nature and win*, Newell (1973) praised experimental psychology for producing clear scientific data on cognitive functions. However, he also criticized the field for lacking a way to unify the data. His solution was to create cognitive architectures (Newell 1973, 1990), which are unified, integrated models of cognition, usually specified as computer code. According to Newell's (1990) system level theory, the neural level implements the (micro) cognitive level, which is described by the (micro) cognitive architecture. The level above the (micro) cognitive level is the knowledge level. The knowledge level is unconstrained by the (micro) cognitive architecture, except for the provision that errors will occur if the processing capabilities of the (micro) cognitive architecture are exceeded (Newell 1990).

Micro cognitive architectures, such as ACT-R (Anderson and Lebiere 1998), SOAR (Laird 2012), and EPIC (Kieras and Meyer 1997), have been successfully used to model macro level tasks, demonstrating that the cognitive mechanisms derived from experimental psychology can scale up beyond the experimental paradigms that produced them (see West and Nagy 2007). However, it is possible to model the same high level task in significantly different ways using the same micro cognitive architecture by changing the knowledge entered into the model (Cooper 2007). So, although these models demonstrate that existing cognitive functions can model macro level tasks, micro cognitive architectures put very few constraints on the model and provide little guidance other than avoiding overload.

The macro architecture hypothesis proposes that Newell's system level scheme be modified to include a macro systems level in between the micro level and the knowledge level (see Fig. 1). Like the micro cognitive architecture, the macro cognitive architecture is proposed to be more or less constant across individuals and across tasks. However, this does not mean that everyone will use the same strategies and procedures. Instead, it means that everyone will use the same general system for managing and integrating different strategies, as well as for other common macro level functions such as dealing with interruptions, re-planning, sense making, and coordinating with others.

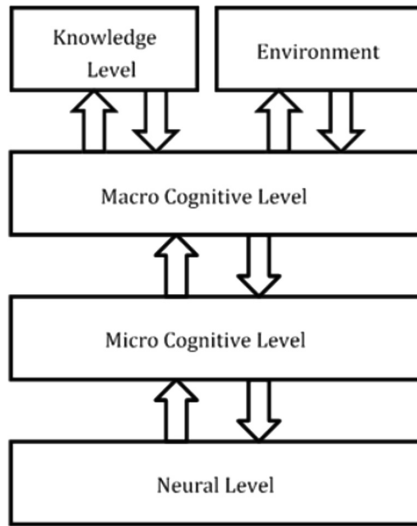


Fig. 1. Modified system levels accommodating a macro cognitive architecture (MacDougall et al. 2014)

To illustrate the difference between a macro architecture and a micro architecture we can examine the relationship between the SGOMS (macro) and ACT-R (micro) architectures. ACT-R describes human cognition in terms of parallel modules, each responsible for an element of cognition (procedural memory, declarative memory, motor system, visual system, etc.). Although parallel, the system is driven by production rules, stored in procedural memory, that fire serially and coordinate the parallel activities of the modules.

Just as micro cognitive architectures are ultimately meant to be built on neural architectures, macro cognitive architectures are meant to be built on micro cognitive architectures. In our case, we have implemented SGOMS (described below) in ACT-R (see West and Pronovost 2009; Somers and West 2013). As noted above, there are many different ways a macro level task could be modelled in a micro cognitive architecture. SGOMS applied to ACT-R, provides a systematic way to model macro tasks in ACT-R (see Ritter et al. 2006, for a review of other attempts at systematizing model building in micro architectures). To do this the basic constructs and functions of SGOMS are deconstructed and built in ACT-R. SGOMS can stand on its own, without ACT-R, or it could be implemented in a different micro cognitive architecture, which would change some of the bottom up constraints on SGOMS (for example, EPIC allows productions to fire in parallel).

Existing ACT-R models related to aspects of macro cognition can be treated as modules within the SGOMS architecture. However, by modules we do not mean dedicated brain areas. Rather we mean set ways of integrating different brain areas to produce specific higher level functions (e.g., Varela et al. 2001). For example, the ACT-R/SGOMS architecture incorporates Salvucci and Taatgen's (2008) ACT-R model of multitasking. The SGOMS architecture also makes use of the ACT-R model

of instance-based reasoning (see Thomson et al. 2015) for high level heuristic decision making.

The macro architecture hypothesis represents a way of systematizing the relationship between micro and macro cognition, with a clear systems level where the study of macro cognition fits in (see Fig. 1). It is also a hypothesis as it makes two important claims: (1) there is a macro cognitive architecture that is relatively invariant across people and across tasks, and (2) it is meaningfully constrained and fully accounted for by micro cognition. If there really is a macro cognitive architecture, then it has major implications for systems design. Specifically, task structures that fit naturally with the architecture should be easier to learn and less error prone, whereas tasks structures that go against the architecture should be harder to learn and more error prone.

1.1 Unit Tasks

Although Newell did not have a separate systems level to mediate between (micro) cognition and real world tasks, he did have a control mechanism. The *unit task* was hypothesized to mediate between the structure of the task and the abilities of the (micro) cognitive system (Newell 1990). Specifically, the unit task defined how the task was mentally broken up to avoid both overloading the cognitive system and down time. For example, a task that involves remembering would be broken down into parts so that the capacity of short-term memory would not be overloaded. Likewise, parts of the task that will not necessarily follow each other would be stored as separate unit tasks, so that the agent can be released in between to do other things if there is time (i.e., avoid downtime).

In cognitive modelling, it is common practice to first determine the unit tasks used in a task, then to model each unit task and some sort of system for choosing which unit task to use next (e.g., see Gray et al. 1993). Models built in this way can be viewed as implicitly embodying two hierarchically arranged systems - the processes contained within the unit tasks and the system for selecting and coordinating the unit tasks. However, most psychology experiments fit within a single unit task. Likewise, most applied modelling projects examine only part of the project, usually corresponding to one or a few unit tasks. Consequently, there has been very little work on the system for selecting which unit task to do next.

1.2 SGOMS

SGOMS is designed to model expert behaviour. The dominant approach in the study and modeling of expertise is to treat each domain of expertise separately (Ericsson et al. 2006; Kirlik 2012). In contrast, SGOMS assumes that expertise is based on a macro architecture that is relatively invariant across different types of experts.

SGOMS is an extension of GOMS (Card et al. 1983). GOMS analyzes tasks in terms of the agents *Goals* and the motor and perceptual *Operators* that the agent uses to accomplish their goals. Frequently repeated strings of operators are represented as *Methods* and *Selection Rules* are used to choose which method or operator to do next. Similar to ACT-R, SOAR, and EPIC, the selection rules are production rules. GOMS is a family of modelling systems that follow the GOMS principles (see John and Kieras

1996). GOMS models can also be implemented in ACT-R, SOAR, or EPIC. The main limitation of GOMS compared to other architectures is that it assumes a task is well learned and does not account for learning.

SGOMS was created when we found that GOMS was unable to handle the frequent interruptions, task switching, and re-planning in real world tasks (see Kieras and Santoro 2004; West and Nagy 2007). To fix this, we modified the definition of the unit task by adding the criterion that a unit task should be small enough so that it will most likely not be interrupted. That is, we defined the unit task as a control structure that functions to avoid overload, downtime, *and interruptions*. This modification allows the unit task to continue to serve its original function to define islands of work that can be executed in a well-defined way.

We also added a second control structure, called the *planning unit*. In SGOMS, the unit task mediates between the micro cognitive level and the macro cognitive level, while the planning unit mediates between the macro cognitive level and the real world, as represented in our perceptions and knowledge of it. In contrast to unit tasks, planning units are designed for interruptions and task switching. Planning units also allow efficient communication and coordination between agents by functioning as the building blocks for creating plans and modifying them. For example, planning units are theorized to have names that are used in communication to establish common ground (Klein 2004) between agents.

The simplest form of planning unit is an ordered list of unit tasks. If a planning unit is interrupted, the current unit task is either finished or abandoned and the situation is assessed. The task can be resumed, or a new planning unit can be chosen based on the current constraints. When a planning unit is interrupted, progress on the planning unit is stored in memory so that it can be resumed, and a new planning unit is chosen. The highest level of decision-making is choosing which planning unit to work on based on the current context, which is constantly updated during the execution of the task. If there is a plan, then that is also part of the context. In addition, each planning unit is associated with a set of constraints.

Planning unit choice is based on either memorized rules (for fast, emergency situations), or memory-based heuristics (for slower, more complex decisions, see West and Nagy 2007, for an SGOMS example). Both of these can be modelled in ACT-R (see Thomson et al. 2015, for a discussion of using ACT-R to model heuristics). SGOMS does not specify what heuristics should be used, this is up to the modeler, instead SGOMS is a system for managing this process. This involves coordinating: (1) low level and parallel functions, such as bottom up and top down perceptual and motor actions, (2) expert knowledge representations, such as production rules (representing procedural memory) and expert knowledge (represented in declarative memory), (3) specific plans for coordinating agents, (4) updating and maintaining representations of all the factors and parameters relative to the task (i.e., context or situation awareness), and (5) Using heuristics, knowledge of the task and the context to re-plan and adjust to unexpected events.

SGOMS has the following hierarchical structure of representations. Each is associated with a different set of cognitive mechanisms:

- Rules and heuristics for selecting planning units based on context
- Planning units - sets of unit tasks to execute, can be interrupted and re-started
- Unit tasks - expert systems for choosing methods, smart but brittle
- Methods - fixed set of actions, executed ballistically
- Operators - basic units of perceptual and motor actions
- Bottom up monitoring - when not busy with top down commands, the system checks the environment and memory for relevant information

The level above controls the level below, but the resources are shared. So, for example, different planning units can call on the same unit task. Figure 2 shows the cycle of operations. Interruptions can occur at any level and may be solved on any level. For example, unit tasks can solve expected or common interruptions related to that unit task because it is part of the routine process. Only if an interruption percolates to the top does it result in re-planning.

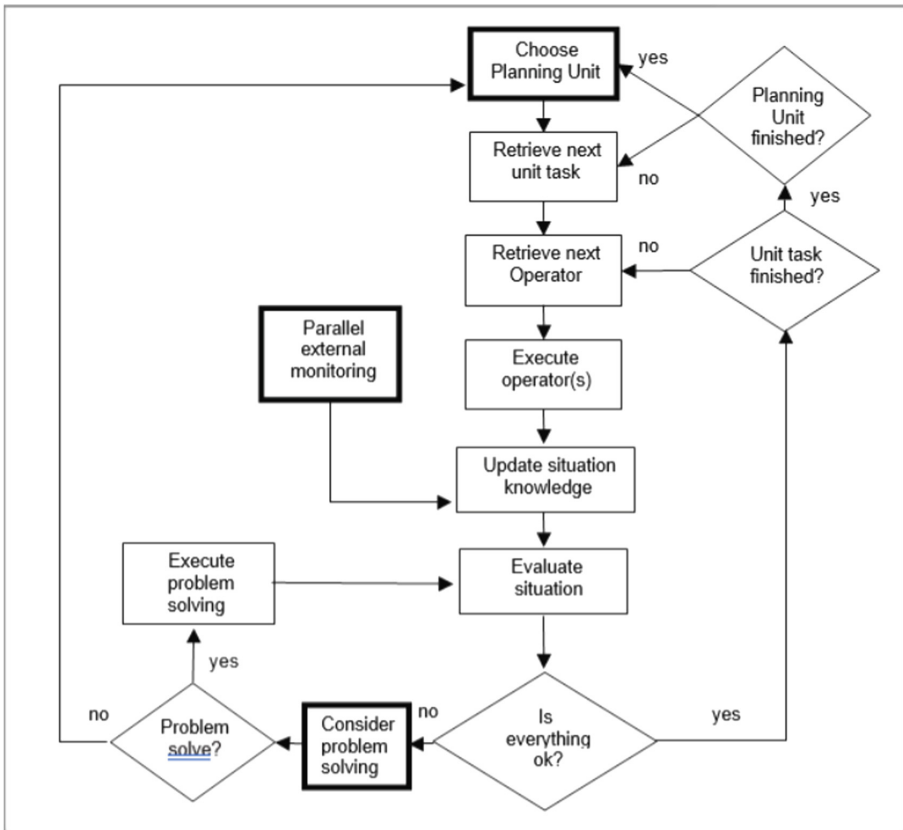


Fig. 2. SGOMS cycle of operations (West and Nagy 2007)

1.3 Model Development

SGOMS provides a way of analyzing the macro portion of a task in terms of planning units and unit tasks. To get to the micro level the model is implemented in a micro cognitive architecture, ACT-R in this case. For this, SGOMS imposes a specific way of modelling within ACT-R, which consists of a set of: (1) task generic production rules for managing planning units, dealing with interruptions, and updating context, (2) task generic ways of representing information in declarative memory (required to interact with the generic production rules), and (3) a set of goal buffers instead of just one (similar to Salvucci and Taatgan 2008). An SGOMS analysis of the planning units and unit tasks can be entered directly into this framework. To make it into a functioning micro cognitive model, the unit tasks and the rules and heuristics for selecting planning units must be modelled in detail. Also, an appropriate model of the task environment must be created (we use Python ACT-R for this, see Stewart and West 2006). Implementing SGOMS in ACT-R puts major constraints on SGOMS and also provides some difficult challenges for ACT-R.

So far, we have tested the SGOMS architecture on telecommunication network maintenance workers (West and Nagy 2007), video game team play (Pronovost and West 2008a, 2008b; West et al. 2013), aviation (Somers and West 2012, 2013) and professional mediation for disputes (West et al. 2013). For mediation, video games, and network maintenance, model tracing showed that there were no cases where the model could not reasonably predict the human behaviour, although in some cases some minor adjustments to productions rules were required.

More generally, in terms of evaluating the SGOMS architecture, or any other macro architecture, the key to is to show that it works across the all examples of the class of behaviours it is supposed to model. SGOMS is meant to model expertise, therefore, SGOMS should work across all forms of expertise. As noted above, this is contrary to current practices and theory concerning expertise, as they treat each area of expertise separately.

2 Experiment 1: Alphabet Expert Task

Within Experiment 1, we analyzed the two main ways that planning units are organized in SGOMS. The first is an ordered list of unit tasks and the second is a set of unit tasks that are cued by the environment. These are known as ordered planning units and situated planning units respectively. The Alphabet Expert task, which is described below, was a speeded stimulus-response task. Subjects were taught three different stimulus response patterns, which represented unit tasks in the SGOMS system. Then subjects were told three separate ways to order the unit tasks. The orders represented planning units in the SGOMS system. However, without imposing the SGOMS structure, these instructions can simply be viewed as describing a hierarchically organized task.

We used the SGOMS template to create an SGOMS ACT-R cognitive model of this task. We also created an optimal ACT-R cognitive model of the task. As predicted, the SGOMS model was slower than the optimal model on specific parts of the task due to

the overhead produced by keeping track of where it was in terms of planning units and unit tasks. This feature of SGOMS is required in order to tolerate interruptions and allow for re-planning. We predicted that subjects would take an SGOMS approach, as it is better for real world, macro level tasks, where interruptions and re-planning are common.

The concept of methods relates to a long history in human factors (Meyer and Kieras 1997). However, the term “methods” is best known as a part of the GOMS modelling system (Card et al. 1983). In the GOMS modelling system a method is a way to achieve a specific sub-goal in the task. The GOMS approach assumes that cognitive, perceptual, and motor actions can be described as distinct, independent operators. Operators describe different actions within a task, such as recall target, move hand to mouse, move eyes to icon, move cursor to icon, and click mouse. Methods are usually ways of organizing operators to achieve sub goals. For example, the chain of actions described above could be considered a method for clicking on an icon. Methods are specific to the interface and the task. They are learned and are assumed to be reused. For example, it is generally assumed that people use the same method for clicking an icon each time (parameterized to suit each instance by considering factors such as distance, target size, etc.). Therefore, for experienced users operating simple interfaces, there will only be a limited number of methods that could reasonably be used (Newell 1973; Card et al. 1983; Gray and Boehm-Davis 2000).

In CPM GOMS (John and Kieras 1996), different operators can be used in parallel to accomplish goals. Usually these models are constructed in the form of a PERT chart where the term, *templates*, is used to describe common ways of organizing and interleaving operators (Gray, John, and Atwood, 1993; John and Kieras 1996). Gray et al. (2000) also used the term, *micro strategies*, to describe what appears to be the same thing as templates. However, as Vera et al. (2005) point out, Gray and Boehm-Davis’s (2000) work on this concept elevates it from a descriptive tool in CPM GOMS (templates), to an actual theory of how the cognitive system interacts with the environment (micro strategies). Vera et al. (2005) also suggest using smaller units, called Architectural Process Cascades, for describing these interactions.

We will use the term *micro strategies*, to refer to low-level strategy decisions for completing a task. However, in this study we were interested in *perceptual/motor* micro strategies only for purposes of controlling for them. Our main purpose was to see if we could detect the influences of *cognitive* micro strategies. By cognitive micro strategies we mean low-level strategies related to the internal processing of information. Our goal in this study was to use models to predict differences in cognitive micro strategies at specific points in the task and then test for these differences in human subjects.

2.1 Procedure

The key elements for this type of experiment are (1) having a very simple response pattern so that the perceptual/motor micro strategies can be isolated, (2) having pre-existing, models representing contrasting options for understanding the task, (3) a highly detailed, model driven analysis of the results, and (4) an analysis based on the results of individual subjects.

2.2 Subjects

Two subjects were analyzed within the Alphabet Expert task. Two of the authors, NN and FK volunteered. Neither had experience with the SGOMS ACT-R model at the time of testing. FK was an experienced video game player, while NN was not.

2.3 Method

To account for variations in methods and unit tasks between participants, we kept our experimental task as simple as possible. To this end, we created a task called the Alphabet Expert, designed to limit response method variability and produce clear unit task structures. Each trial, subjects were presented with a four-letter code and were required to respond with the appropriate, corresponding two-letter code. Therefore, each trial was identical in terms of required response actions. However, trials were designed to include sequences in which participants knew which prompt code would occur in a predictable order, while other sequences were randomized, therefore required participants to perceive the prompt code to know which code to respond with. In this study, we attempted to find evidence for the SGOMS architecture. To do this we adopted the approach used in (Gray and Boehm-Davis 2000) to study micro strategies, replicating the two conditions of the experiment within a game-play task.

2.4 Training

Phase 1: Subjects were required to learn three distinct unit tasks, which were presented individually as sequential units. Unit tasks were presented with one-second intervals between prompt code presentation and response code. Subjects trained until they had attained their best speed and accuracy of unit task performance. To do this, subjects trained at home on their personal mobile phones. Figure 3 illustrates the structure of each unit task.

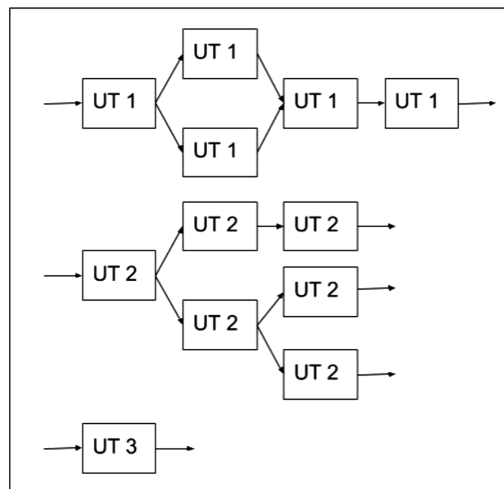


Fig. 3. Unit task structures.

Phase 2: Subjects were required to learn three distinct planning units separately. The start of each planning unit began with a distinct prompt cue consisting of a four-letter code, followed by a sequence of three unit tasks (introduced in Phase 1). Two of the planning units were ordered planning units, as they were designed to consistently present the same unit tasks in a planning unit specific order. The third planning unit learned was a situated planning unit, as each time the three unit tasks were presented in an unpredictable and random order. The planning unit structures are presented within Table 1. As training Phase 2 included a prompt code prior to planning unit commencement, participants were aware of which planning unit they were practicing each time.

Table 1. Planning unit structure.

| Planning unit | Unit task order |
|-------------------------|-----------------|
| Ordered planning unit 1 | Unit task 1 |
| | Unit task 2 |
| | Unit task 3 |
| Ordered planning unit 2 | Unit task 2 |
| | Unit task 3 |
| | Unit task 1 |
| Situated planning unit | Unit task ? |
| | Unit task ? |
| | Unit task ? |

Phase 3: Subjects then practiced the whole task with one-second intervals between trials. For the ordered planning units, when subjects were cued with the code for the planning unit, they needed to recall the first unit task and type in a code to begin it. When that unit task was finished, the code for the planning unit was presented again and subjects had to type in a code for the next unit task. The sequence continued until the planning unit was finished. For the situated planning unit, subjects were cued with a code specific to indicate the situated planning unit. Subjects responded by inputting a code to start the situated planning unit. Then subjects were cued by the code for the beginning of a unit task. After completing the unit task, subjects were presented with the same prompt code indicating continuation of the situated planning unit, requiring the same appropriate response to continue. As before, the cue for the next unit task was presented. This pattern was then repeated a third time. The order of three unit tasks presented within the situated planning unit was always random. Subjects practiced on their personal devices at home until they were satisfied they were doing the task as quickly and accurately as possible.

Phase 4: For the final training stage, the one second interval between trials was removed, so as not to mask cognitive actions, and subjects once again practiced until they were satisfied they were going as quickly and accurately as possible. Once this stage was reached, we collected the data for analysis.

2.5 Model Development

Two models were developed to predict the human behaviour in this task: an SGOMS model that applied the same knowledge structures that were practiced by participants, and a model optimized for speed to produce the fastest possible game play using ACT-R. We used perceptual/motor method time estimates across all conditions for both models. With the methods determined, the only difference between the two models was the number of productions, which were determined before the experiment began. The SGOMS ACT-R model was created by writing the code for the unit tasks, the planning units and the perceptual/motor methods, and inserting them into the SGOMS ACT-R template. The optimized model worked in the following way: the code for the current planning unit was stored in the imaginal buffer (i.e., working memory) and the production representing the correct response was selected by matching with this information and the current code, which was in the visual buffer. The perceptual/motor methods were the same as within the SGOMS knowledge model.

2.6 Alphabet Expert Results

To evaluate the results, we divided the trials into distinct categories that corresponded to different predictions of the SGOMS architecture, represented in Table 2. In SGOMS, an action occurring inside a unit task occurs as it would in the optimal ACT-R model. That is, there is no overhead. These response categories were labeled Unknown Unit Task Middle (Unknown Mid UT) and Known Unit Task Middle (Known Mid UT), where Known refers to conditions in which the subject knew the next response based on the last response, and Unknown refers to conditions in which the subject had to read the new code to know the right response.

For the known response, we assumed a single perceptual/motor method, where the subject entered the next response as fast as possible. For the unknown response, we assumed the subject used two perceptual/motor methods, the first to identify the code and the second to enter the response. After removing trials with an error and outliers more than two standard deviations from the mean, the reaction times of the two conditions were very consistent, with the unknown condition taking longer, as expected. These two conditions formed the baseline for fitting the results. FK's response times were considerably faster than NN, in part because FK used two hands to type and NN used one, but also possibly because FK was an avid video game player and NN was not. We equalized the response times by subtracting the difference between FK's average RT and NN's average RT from NN's average score for both conditions. We applied this same correction to each condition based upon whether the response was known or unknown. Figure 2 shows the results with 0.05 confidence intervals for our subjects' data. For the response categories related to the ordered planning units, NN and FK were the same and not significantly different from the SGOMS model for all but one response category (Known First Unit Task) where NN matched the SGOMS model, and FK matched the optimal model (Fig. 4).

Table 2. Response category descriptions.

| Response category | Description |
|--------------------|--|
| Unknown Mid UT | A response in the middle of a unit task that is not known until the code is perceived |
| Known Mid UT | A response in the middle of a unit task that is determined by the response before it |
| Unknown PU-O Start | The response to the code to begin an ordered planning unit. This response cannot be determined by the response before it |
| Known PU-O Mid | The response to the code to begin the second or third unit task in an ordered planning unit. This response can be determined by the response before it |
| Known PU-S Mid | The response to the code to begin the second or third unit task in an unordered planning unit. This response can be determined by the response before it |
| Known First UT | The response to the first unit task in an ordered planning unit. This response can be determined by the response before it |
| Unknown First UT | The response to the first unit task in a situated planning unit. This response cannot be determined by the response before it |
| Unknown PU-S Start | The response to the code to begin a situated planning unit. This response cannot be determined by the response before it |

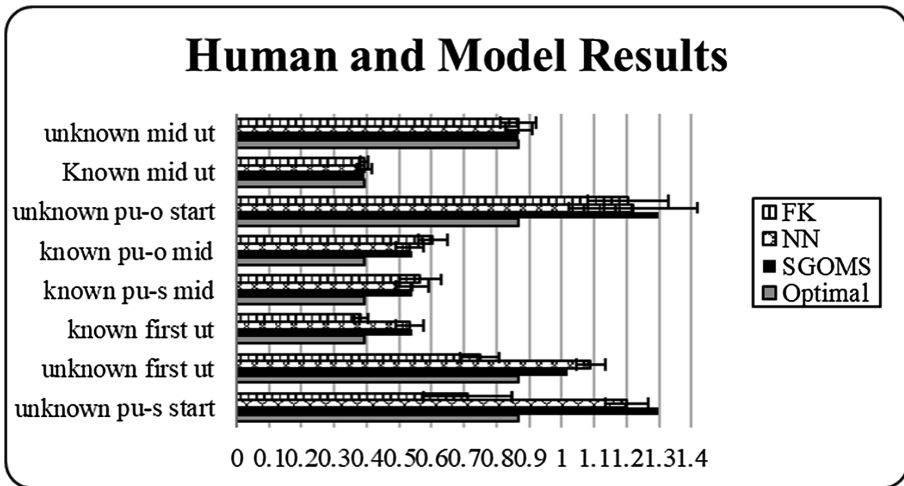


Fig. 4. Human and model results.

For the response categories related to the situated planning units, the differences between NN and FK were greater. NN was similar to the SGOMS model. In two cases, there was a significant difference between NN and the SGOMS model (Unknown First UT and Unknown PU-S Start), however, NN’s results were still closer to the

SGOMS model than the optimal model. FK matched the SGOMS model in some cases, and the optimal model in other cases. In two cases FK was significantly faster than the optimal model by a small amount (again, these were Unknown First UT and Unknown PU-S Start). FK reported experimenting with different strategies to speed up the task. The pattern of FK's results indicated that heightened training lead to the elimination of the SGOMS processes for some parts of the task, as performance matched the optimal model. FK's response times within the situated planning unit and for the first response for each unit task in the situated planning unit were due to his strategies implemented to focus on speeding up this part of the task, therefore achieving faster response times. However, FK's responses to the prompt cue to begin the second and third unit tasks within the situated planning unit were as predicted by the SGOMS model (Known PU-S Mid), indicating that vestiges of the SGOMS process remained.

Based upon self-report data, consistent with FK's background as an avid video game player, we believe that FK was attempting to reorganize his understanding of the task so that actions were faster. NN also reported testing different strategies to improve performance. Both NN and FK focused most of their strategic thinking on the situated planning unit. This is also where the significant differences occurred, which may indicate that situated planning units are a more obvious candidate for optimization. As noted above, we are not claiming that people cannot learn the optimal way to do a task. Our research hypothesis claiming that the optimal way to perform a task is not the default starting point for most real-world tasks is supported by the results of the Alphabet Expert task.

2.7 Alphabet Expert Conclusion

SGOMS knowledge model provided a better fit to the data for six out of six response categories for NN, and for three out of six response categories FK. For FK, in the three cases where the SGOMS model did not match the data, the optimal model provided a reasonable match. The pattern of results supports our claim that people use the SGOMS architecture as their default system, and only later convert to an optimal form if the task can be performed without interruptions and they have the motivation for thinking about it and practicing it. This could potentially be understood in terms of Lebiere and Best's (2009) strategy evolution and strategy discovery levels. More broadly, our results support the idea that macro level factors systematically shape default strategies for using our micro cognitive architecture.

Macro cognitive architectures are descriptions of these regularities and how they are related to dealing with specific classes of real world problems. For example, SGOMS describes how people execute expert knowledge in real world environments, where interruptions and re-planning are common. Given that people rarely take part in psychology experiments, it is not surprising that they would default to a strategy that works well in their daily life. As we have shown, the Gray and Boehm-Davis (2000) methodology works well and can be employed to study the effects of macro cognitive architectures using micro cognitive experimental and modeling methods.

3 EOC Management

In the second study, we examined SGOMS predictions on a very different scale. This study is part of an ongoing project to use SGOMS to model the behaviour of Emergency Operations Centre (EOC) workers during disaster simulations, with the goal of gaining insight into training, policy, and systems design. The following is a working definition of EOCs:

*“An **emergency operations center (EOC)** is a central command and control facility responsible for carrying out the principles of emergency preparedness and emergency management, or disaster management functions at a strategic level during an emergency, and ensuring the continuity of operation of a company, political subdivision or other organization.*

An EOC is responsible for strategic direction and operational decisions and does not normally directly control field assets, instead leaving tactical decisions to lower commands. The common functions of EOCs is to collect, gather and analyze data; make decisions that protect life and property, maintain continuity of the organization, within the scope of applicable laws; and disseminate those decisions to all concerned agencies and individuals.” (Emergency operating center, Wikipedia.)

EOC managers can only make recommendations and pass on information, they cannot force the field commanders to do things. The EOC manager, who is free from ongoing distractions in the field, can maintain a detailed and accurate model of the situation and take the extra time to make strategic recommendations. The recommendations and the reasons for the recommendations are passed onto the field commanders, which frees them from time-consuming data consolidation and provides them with a broader context. However, at the same time the field commanders can see if there is anything locally wrong with the recommendations based on their more immediate and detailed knowledge of the situation on the ground.

EOC management is obviously an important function, where failures have very serious consequences (e.g., consider the response to the Fukushima nuclear disaster). To get an overview of EOC management skills, here is a summary of the abilities an EOC manager should have from the U.S. Federal Emergency Management Agency (FEMA). Note that the descriptions are very vague but mostly refer to high-level decision-making:

- Comprehensive – emergency managers consider and take into account all hazards, all phases, all stakeholders and all impacts relevant to disasters.
- Progressive – emergency managers anticipate future disasters and take preventive and preparatory measures to build disaster-resistant and disaster-resilient communities.
- Risk-driven – emergency managers use sound risk management principles (hazard identification, risk analysis, and impact analysis) in assigning priorities and resources. Integrated – emergency managers ensure unity of effort among all levels of government and all elements of a community.
- Collaborative – emergency managers create and sustain broad and sincere relationships among individuals and organizations to encourage trust, advocate a team atmosphere, build consensus, and facilitate communication.

- Coordinated – emergency managers synchronize the activities of all relevant stakeholders to achieve a common purpose.
- Flexible – emergency managers use creative and innovative approaches in solving disaster challenges.
- Professional – emergency managers value a science and knowledge-based approach; based on education, training, experience, ethical practice, public stewardship and continuous improvement.

Macro level analyses of real world tasks are more similar to anthropology or sociology than to experimental methodology. Cognitive psychology comes into it mainly in terms of a language for describing principles and heuristics for strategic decision making, where the data sources are usually interviews and observations. Using a macro cognitive architecture provides a principled framework for understanding a task, which provides a basis for creating design improvements. In this spirit, we applied SGOMS to understanding EOC management.

3.1 SGOMS and EOC Professionals

For this study we used observations from actual EOC professionals in disaster simulations, as well as inexperienced undergraduates in similar simulations. In the simulations, performance did not depend on being fast, which is what GOMS traditionally focuses on. Instead, success was related to sense making and the ability to support decision-making. The process of choosing planning units within SGOMS provides a good way of understanding this process. In SGOMS, planning units are chosen using constraint-based decision making. This involves understanding the current context, or constraints, and using various heuristics for bounded rationality decision making. The agents in the field must do this for whatever part of the disaster they are working on. So, using the SGOMS structure, we can conceptualize the EOC manager as providing context (constraints) and sometimes suggesting solutions (heuristic based rational analysis) to help the agents in the field choose appropriate planning units. Essentially, the EOC manager functions as a planning unit recommender system.

Planning units are also used as a quick efficient way to communicate and coordinate, but it requires common ground (MacDougall et al. 2014). One prediction that arises from this is that this will function efficiently to the degree that the EOC managers and agents in the field use a common set of planning units. In terms of the professional EOC management teams that we observed, this seemed to play out when observing the difference between teams composed of ambulance, police, and fireman versus teams composed of trained EOC managers. The ambulance, police, and fireman seemed to be more efficient in their reactions. This appeared to be due to their shared labeling system for planning units related to emergency response.

3.2 Model of Novices

Our approach for modelling the undergraduates was to assume that people have a cognitive template for understanding and implementing instructions in a task. We further assumed that this template amounts to using an SGOMS-like structure to

interpret the instructions. Half way through the task and at the end of the task the participants were asked to report all the events. The main finding for the undergraduates was that they forgot a surprising number of events. We focused on that as it indicated a lack of situation awareness, which is critical for EOC management.

The model for the undergraduate participants was relatively simple as their task was limited. Also, because this was not a speeded task, the ability of GOMS to predict completion times was not useful. Participants had more than enough time to complete the components of the task, so minor differences in speed would not impact performance. Therefore, we did not model the low level perceptual motor components of the task.

Another issue was that some components of the task could not be handled by GOMS. Specifically, GOMS is not designed to model sense making or the composing of notes and reports. However, task components that cannot be directly modelled in GOMS can still be represented by unit tasks.

For this model we used unit tasks for sense making, writing, and information gathering. We also used the interrupt mechanism in SGOMS to deal with new incoming information. SGOMS is capable of ignoring irrelevant information but in this case, we assumed, due to the task, that participants would pay attention to all of the incoming information. The model operated in the following way: when new information was received, it triggered an interruption to the current planning unit. The model would then switch to the *new-information* planning unit. This was an ordered planning unit consisting of, first, the *attend-new-information* (ANI) unit task followed by the *select-incident-to-work-on* (SITWO) unit task. The SITWO unit task represents the metacognitive process of selecting what to work on next. For the student volunteers we modelled this using the availability heuristic. This was implemented using the ACT-R declarative memory system, which has been shown to be accurate in predicting human forgetting across a wide range of experiments.

Another unique feature of the model was the use of instances. SGOMS does not need a pre-existing planning unit for each instance of a task component. Instead, SGOMS can use a generic planning unit to generate a specific instance of that planning unit. In this case, when a new event occurred, a planning unit for that event would be generated by attaching the identity of the event to a copy of the generic planning unit for processing the events (i.e., writing reports and issuing advice in this case). Thus, the model could create new planning units to represent different events.

After an interruption, the SITWO unit task would use the availability heuristic to decide what to work on next. This was modeled using the ACT-R declarative memory retrieval mechanism to retrieve the most active memory chunk representing an ongoing event. Running the model predicted that, without rehearsal, the participants would forget to report almost everything. However, rehearsal occurred in the model every time an event was worked on. The key to remembering was to work on an event.

3.3 EOC Simulation Results

Most of the failures to recall an event could be placed into three categories that could all be accounted for with the model. The first was completion errors. These occurred when an event seemed to be completed, in that it was appeared to be no longer a

problem. This can be accounted for in the model by assuming that coding an event as being completed blocks it from being retrieved when the availability heuristic is used. This could be modeled by making one of the retrieval criterions that the event be active (i.e., not completed), or it could be modeled using the spreading activation mechanism in ACT-R. Either way, this would prevent the event from being retrieved and worked on further, allowing the activation level to decay below threshold.

The second category of forgetting occurred when new event information was received during the processing of incoming information from another event. This includes when a new event was presented, or when there was a change in the narrative of a previously presented event (e.g., an ice storm has knocked out power but then causes the roof of an ice rink to collapse). This type of forgetting can be modeled by assuming that when the novice participants were interrupted they did not rehearse or further process the interrupted task before switching from it. This would prevent the interrupted task from getting a memory boost from having been worked on.

The third category of forgetting occurred when two events with nearly identical features (e.g. resource required, emergency type) at different locations were presented within a close time frame in the simulation. In the model, this type of error would be a natural outcome of the ACT-R partial matching mechanism. If a chunk representing one event had a higher activation and was also very similar to a chunk representing another event, the higher activation chunk would always be retrieved, thus making the less active chunk unavailable to the availability heuristic.

3.4 EOC Training Recommendations

Based on using SGOMS to frame our understanding of EOC management, we derived the following recommendations for training. The first is to have a clear labeling system for the planning units used by EOC management and the field assets. Second, these labels need to be practiced, we recommend the use of practice simulations for this. Finally, when training EOC operators there should be a focus on maintaining situation awareness through deliberate memory storage strategies. For example, we noticed that the professional teams used maps to organize themselves, suggesting that they were using location to encode knowledge in memory. This is also a place where better technology could help. For example, having automatic transcription of information from telephone, radio, and TV might help, as these sources leave no trace except in memory.

4 Conclusion

In this paper we showed how SGOMS could be used to model two very different tasks. In the alphabet expert task we showed how SGOMS could model the effects of training and strategy, and make highly accurate reaction time predictions. In the EOC task we showed how SGOMS can be used as a framework to understand high level tasks and to model particular parts of the task, memory in this case. More generally, the ability of SGOMS to usefully model both of these tasks supports our argument that SGOMS is appropriate for modelling any expert task.

References

- Anderson, J.R., Lebiere, C.: *The Atomic Components of Thought*. Erlbaum, Mahwah (1998)
- Cacciabue, P.C., Hollnagel, E.: Simulation of cognition: applications. In: Hoc, J.M., Cacciabue, P., Hollnagel, E. (Eds.), *Expertise and Technology: Issues in Cognition and Human-Computer Cooperation*, pp. 55–74. NEA, Hillsdale (1995)
- Card, S., Moran, T., Newell, A.: *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, Hillsdale (1983)
- Cooper, R.P.: The role of falsification in the development of cognitive architectures: Insights from a lakatosian analysis. *Cogn. Sci.* **31**(2007), 509–533 (2007)
- Emergency Operating Center: Wikipedia online (n.d.). https://en.wikipedia.org/wiki/Emergency_operations_center
- Ericsson, K.A., Charness, N., Hoffman, R.R., Feltovich, P.J. (eds.): *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, New York (2006)
- Federal Emergency Management Agency, Department of Homeland Security: *Exercise Simulation System Document [ESSD]* (2014). <https://training.fema.gov/fiemc/exercisesimulationdocument.aspx>
- Gray, W.D., Boehm-davis, D.A.: Milliseconds matter: an introduction to microstrategies and to their use in describing and predicting interactive behavior. *J. Exp. Psychol.* **6**(4), 322–335 (2000)
- Gray, W.D., John, B.E., Atwood, M.E.: Project Ernestine: validating a GOMS analysis for predicting and explaining real-world task performance. *Hum. Comput. Interact.* **8**(3), 237–309 (1993)
- Gregson, R.A.M.: *Nonlinear Psychophysical Dynamics*. Erlbaum Associates, Hillsdale (1988)
- John, B.E., Kieras, D.E.: The GOMS family of user interface analysis techniques: comparison and contrast. *ACM Trans. Comput. Hum. Interact.* **3**(4), 320–351 (1996)
- Kieras, D.E., Meyer, D.E.: An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *J. Hum. Comput. Interact.* **12**(4), 391–438 (1997)
- Kieras, D., Santoro, P.: Computational GOMS modeling of a complex team task: lessons learned. In: *Proceedings of HCI, 24–29 Apr 2004, Vienna, Austria*, pp. 97–104 (2004)
- Kingstone, A., Smilek, D., Ristic, J., Friesen, C.K., John, D., Eastwood, J.D.: Attention, researchers! It is time to take a look at the real world. *Curr. Dir. Psychol. Sci.* **12**, 176 (2003)
- Kirlik, A.: The emerging toolbox of cognitive engineering models. In: *Paper presented at the International Conference on Cognitive Modeling, Berlin, Germany, 13–15 Apr 2012*
- Klein, G., Ross, K.G., Moon, B.M., Klein, D.E., Hoffman, R.R., Hollnagel, E.: *Macro-cognition*. *IEEE Intell. Syst.* **18**(3), 81–85 (2003)
- Klein, G., Woods, D.D., Bradshaw, J.D., Hoffman, R.R., Feltovich, P.J.: Ten challenges for making automation a “team player” in joint human-agent activity. *IEEE Intell. Syst.* **19**, 91–95 (2004)
- Laird, J.E.: *The SOAR Cognitive Architecture*. MIT Press, Lakatos (2012)
- Lebiere, C., Best, B.J.: From micro-cognition to macro-cognition: architectural support for adversarial behavior. *J. Cogn. Eng. Decis. Making* **3**(2), 176–193 (2009)
- MacDougall, W.K., West, R., Hancock, E.: Modeling multi-agent chaos: killing aliens and managing difficult people. In: *36th Annual Meeting of the Cognitive Science Society*, pp. 2603–2608 (2014)
- Meyer, D.E., Kieras, D.E.: A computational theory of executive control processes and human multiple-task processes and human multiple-task performance: Part I. Basic mechanisms. *Psychol. Rev.* **104**, 3–65 (1997)

- Newell, A.: You can't play 20 questions with nature and win: projective comments on the papers of this symposium. In: *Visual Information Processing*. Academic Press (1973)
- Newell, A.: *Unified Theories of Cognition*. Harvard University Press, Cambridge (1990)
- Pronovost, S., West, R.L.: A GOMS model of virtual sociotechnical systems: using video games to build cognitive models. In: *Proceedings of the European Conference on Cognitive Ergonomics* (2008a)
- Pronovost, S., West, R.L.: Bridging cognitive modeling and model-based evaluation: extending GOMS to model virtual sociotechnical systems and strategic activities. In: *Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society* (2008b)
- Ritter, F.E., Haynes, S.R., Cohen, M.A., Howes, A., John, B.E., Best, B., Lebiere, C., Jones, R. M., Lewis, R.L., St. Amant, R., McBride, S.P., Urbas, L., Leuchter, S., Vera, A.: High-level behavior representation languages revisited. In: *Proceedings of the International Conference on Cognitive Modeling*, Trieste, Italy, pp. 404–407 (2006)
- Salvucci, D.D., Taatgen, N.A.: Threaded cognition: an integrated theory of concurrent multitasking. *Psychol. Rev.* **115**, 101 (2008)
- Stewart, T.C., West, R.L.: Deconstructing ACT-R. In: *International conference on cognitive modelling*, Trieste, Italy (2006)
- Somers, S., West, R.L.: Macro cognition: using SGOMS to pilot a flight simulator. In: *Proceedings of the Annual ACT-R Workshop* (2012)
- Somers, S., West, R.L.: Steering control in a flight simulator using ACT-R. In: *Proceedings of the International Conference on Cognitive Modeling* (2013)
- Thomson, R., Lebiere, C., Anderson, J.R., Staszewski, J.: A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *J. Appl. Res. Mem. Cogn.* **4**, 180–190 (2015)
- Turvey, M.T., Carello, C.: On intelligence from first principles: Guidelines for inquiry into the hypothesis of physical intelligence (PI). *Ecol. Psychol.* **24**(1), 3–32 (2012)
- Van Gelder, T., Port, R.F.: It's about time: an overview of the dynamical approach to cognition. In: Port, R.F., Van Gelder, T. (eds.) *Mind as motion: Explorations in the dynamics of cognition*. MIT Press, Cambridge (1995)
- Varela, F., Lachaux, J.P., Rodriguez, E., Martinerie, J.: The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* **2**(4), 229–239 (2001)
- Vera, A.H., Tollinger, I., Eng, K., Lewis, R., Howes, A.: Architectural building blocks as the locus of adaptive behavior selection. *Proc. Cogn. Sci. Soc.* **27** (2005)
- West, R.L., Nagy, G.: Using GOMS for modeling routine tasks within complex sociotechnical systems: connecting macrocognitive models to microcognition. *J. Cogn. Eng. Decis. Mak.* **1**, 186–211 (2007)
- West, R.L., Pronovost, S.: Modeling SGOMS in ACT-R: linking macro- and microcognition. *J. Cogn. Eng. Decis. Mak.* **3**(2), 194–207 (2009)
- West, R.L., Somers, S.: Scaling up from micro cognition to macro cognition: using SGOMS to build macro cognitive models of sociotechnical work in ACT-R. In: *The proceedings of Cognitive Science*, pp. 1788–1793. Boston, Mass: Cognitive Science (2011)
- West, R.L., Hancock, E., Somers, S., MacDougall, K., Jeanson, F.: The macro-architecture hypothesis: applications to modeling teamwork, conflict resolution, and literary analysis. In: *Proceedings of the International Conference for Cognitive Modeling* (2013)



Natural Interaction in Video Image Investigation and Its Evaluation

Yan Zheng^{1,2} and Guozhen Zhao^{1,2}(✉)

¹ CAS Key Laboratory of Behavioral Science, Institute of Psychology,
Chinese Academy of Sciences, Beijing 100101, China

{zhengy, zhaogz}@psych.ac.cn

² Department of Psychology, University of Chinese Academy of Sciences,
Beijing 100049, China

Abstract. Video image investigation has become an important way to protect the public security. Traditional paper-based investigation method lacks of natural interactive tools to organize case and clue information, leading to poor work efficiency and performance. By interviewing professional video analysts, the current study developed the mental model of the video image investigation and designed an interactive video image investigation system based on the mental model. To evaluate the naturalness and the efficiency of the newly designed system, a user study were conducted involving 28 volunteers. Half of them used the interactive video image investigation system, while the other half used traditional paper-based method to complete an investigation task with a set of real surveillance videos. Through the analysis of brain activity and behavioral indices, we found that the proposed interactive video image investigation system led to higher work efficiency, lower mental workload, and more positive emotional experience or approach motivational tendencies compared to traditional paper-based method. These results suggested the potential application of efficiency, mental workload, and emotional experience for the evaluation of the naturalness of an interactive system.

Keywords: Video image investigation · Natural interaction · Mental model
Mental workload · Emotional experience · EEG

1 Introduction

Surveillance videos contain rich information and are ubiquitous in our life. They play the important role for the public security (e.g., evidence recording, case analyses, crime prevention). Surveillance videos have the typical spatio-temporal characteristics with the two-dimensional property of time and space. The present surveillance video investigation mainly relies on an analyst's scanning, monitoring, judgment, and reasoning to discover clues related to the case. To restore the process of a real case and track the closely related information of the case, a large number of surveillance videos with extremely long video footages must be collected, resulting in a laborious and time-consuming process of video analysis. A large number of surveillance video data also lead to poor performance (e.g., misjudgment, missed sanction), lack of

concentration, high mental workload, fatigue, and bad mood (e.g., frustration, anxiety). Therefore, an interactive system that can facilitate the process of video analysis and improve the job performance and emotional experience is needed.

Paper-based video image investigation is one of the most widely used methods in practical work. Analysts have to save a screenshot when they find any relevant information of the case, and write down all information (e.g., time, location, criminal characteristics) on the paper for the further analysis. Due to individual differences in hands-on background, paper-based investigation information is easy to be interrupted without timely data storage and difficult to be recognized by other analysts. Moreover, clues are usually scattered in different parts of a video or different videos. Analysts need to integrate many video and image fragments of information into a chain of clues to analyze the whole process of the crime. The existing paper-based video image investigation approach is far from our expected goals. Therefore, how to effectively extract the clues from a large number of surveillance videos, simplify the process of video saving, organize the relationship between different clues, and improve the efficiency of the video information retrieval are the interests of our current study. In addition, how to develop a user-centric video image investigation system with an effective and natural way to interact and how to evaluate the naturalness and the efficiency of the system interaction are the most challenging parts of this study.

2 Natural Interaction in Video Image Investigation

2.1 Natural Interaction

Hand-drawn sketch is a kind of natural and direct human thinking externalization and communication mode. The sketch is able to use simple shapes to express the abstract thinking intention. It has the semantic features of texts and images, so when people see a rough sketch, they can immediately know the semantics behind it. In addition, sketch can be used as a simple gesture to interact, such as the abstract, symbolic and fuzzy characteristics. Its simple, rapid and optional features can also serve as a good medium of information expression. Therefore, it is the key research direction of sketch-based analysis of surveillance videos to extract and organize important clues from surveillance videos.

How to develop a user-centric video image investigation system with a natural way to interact? Researchers have made their efforts in interviewing with actual users and recording their subjective thoughts and feelings. In this study, we developed a mental model of the video image investigation to help us to understand surveillance video analysts and designed an interactive video image investigation system for them.

2.2 Mental Model of the Video Image Investigation

Mental model is an internal symbolic representation for actual users in the external world which helps individual to understand, interpret and anticipate how things work [1]. It is dynamic and can be manipulated in mind to obtain outcomes [2]. Mental model has been applied in a variety of product design (e.g., domestic appliances, traffic

and military facilities) to help designers to understand users and make better products for them [3–6]. The most commonly used method is the interview, especially in the development of mental models for the practical applications. There are two categories of interview: structural interview and semi-structural interview. Structural interview implies previously settled questions before interview implement, while semi-structural interview is more flexible and diverse among interviewees. Through an interview, mental models of variable products (e.g., flush toilet, home heating, electronic healthy recording system [7–9]) have been developed.

In this study, to obtain a mental model of the video image investigation, 27 target users (i.e., surveillance video analysts) were interviewed by two experienced experimenters. There were two sessions during the interview. Their demographic information, working history and experience, daily workload, etc. were collected to control individual differences in the development of mental models. In the first session, each video analyst talked about their daily work, for example, what they usually look for in a video to find a suspect, how they deal with different video clips and clue information, and the common contents they write down or save. Two real cases were replayed in which each video analyst was required to explain the detailed investigation procedure step by step. One simple case was analyzed with 10-h surveillance videos from 5 camera locations, while the other one case was more complicated and was analyzed with hundreds-of-hours surveillance videos from more than 20 camera locations.

In the second session of our interview, all video analysts were gathered for a group interview in which they were required to complete a real case together. Before the group interview, all video analysts were informed about what kind of case happened, time and place of the case, and target suspect and vehicles. Their primary work was to watch each surveillance video, mark on the target suspect when it appeared in the video, and draw a road map of the target suspect from the first time appeared in the video to the time and place of the case. They were required to work according to their usual mode of work and use a common video player to play each video and a notebook to write down any relevant clue information. During the task, video analysts were interrupted at each necessary step and were asked to explain their operations and discuss what were the problems at the moment and possible solutions they expected to solve the problems. Here, we listed three major problems reflected in both sessions of the interview:

- Some video clips (e.g., from personal cameras of small merchants) were poorly uniformed and cannot be played by a common video player.
- Important clues were scattered in the notebook with different marking/symbolic styles, which were difficult to be recognized and used by other analysts.
- Clue information retrieval was troublesome because there was no hyperlink between video clips and important clues. Videos must be found out again and played back to the time point based on the information on the notebook.

Based on the in-depth interview with actual video analysts, we drew a mental model (see Fig. 1) to reflect how the video image investigation system should work and interact with video analysts. There were four major components of the video image investigation system: video database, video player, material warehouse, and case management. Case management component was used to create a new case or close an

existing case, assign tasks to different video analysts, and organize all case information by combining and retrieving text and image clues. Video database component and video player component were used to process video clips. Nonstandard video clips were first transformed into a standard format. A standard video clip was selected and played and its time and location information were automatically marked on the road map. When video analysts found a relevant clue, the annotation function was activated to generate a hyperlink connecting a video segment (e.g., segment length can be customized) with the annotation by individual analysts. When a number of videos were investigated, all hyperlinks with the annotation were connected to form a road map that illustrated the suspect’s spatio-temporal activity trajectory. Except for surveillance video, other types of evidence (e.g., images, testimony) were stored in the material warehouse component. In Fig. 1, green parts represented storage space, surveillance videos with hyperlinks, target images and corresponding notes, electric evidence, testimony were stored in two components. Blue parts represented a video player’s major operations and functions. Black parts represented information management and organization during the process of video image investigation.

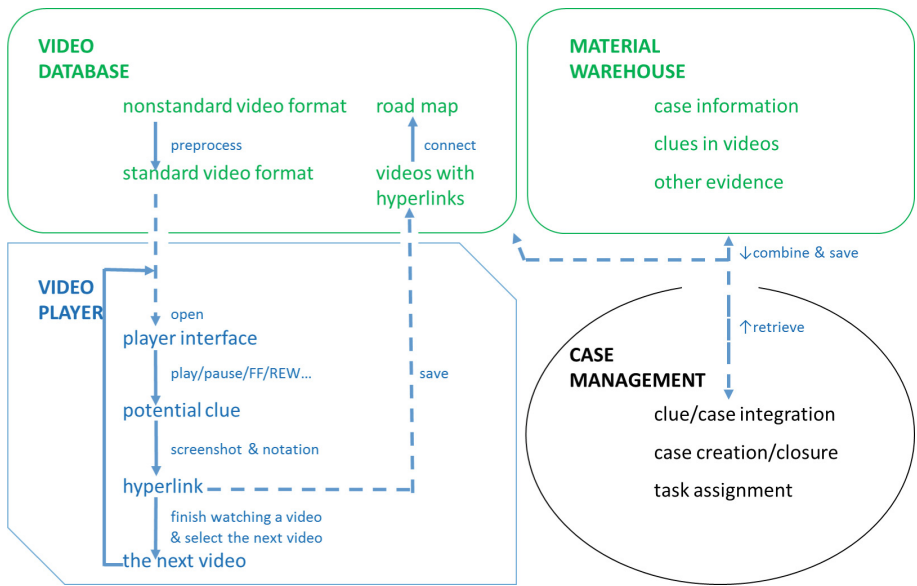


Fig. 1. Mental model of the video image investigation consisted of four components: video database, video player, material warehouse, and case management. Arrows indicated how videos and clue information were passed between four components. (Color figure online)

2.3 Major Functions of the Proposed Video Image Investigation System

According to the mental model of the video image investigation, we designed a new video image investigation system for video analysts to facilitate the process of video

analysis and enhance their job performance and emotional experience. Due to page limitations, we only described two major features of the proposed system.

Sketch-Based Video Annotation. The proposed system uses the combination of sketches and texts to generate video annotations chosen from our sketch tag library. Video annotations includes the time point at which an annotation is added, annotation tag, and the current analyst's information. Specifically, video annotation is divided into two categories: video frame annotation and video clip annotation. Video frame annotation is added when a video suspends and is used to annotate an object or event in different video frames. Video clip annotation is added when a video is playing and is used to annotate a video within a period of time. We take the advantage of simple, intuitive characteristics of sketch to achieve effective content annotation through the annotation of suspects and vehicles, and use the text comments to describe their attributes and characteristics. We use the frame image for video frame annotation and the key frame extracted from the video for video clip annotation. The key frame extraction algorithm is based on clustering, which is on the basis of the similarity between two frames' color histograms. Compared to traditional paper-based annotation, sketch-based video annotation was expected to improve an analyst's work efficiency and the collaboration between multiple analysts on the basis of standard annotation principles.

Hyperlink. When a video annotation is created, a hyperlink is automatically generated connecting the video segment with the corresponding video annotation. There are two major contributions of the hyperlink during the process of video analysis. First, hyperlink is quite useful when analysts tend to play back a video to browse the clues. Traditional method is to find out the target video and navigate to the time point based on the recording of screenshots or notes. In contrast, the proposed system can track back to the target video segment with detailed video annotations using the hyperlink function in a more efficient way. Second, surveillance videos are usually kept for three months, and only those relevant to the important cases are kept for no longer than six months. The proposed system provides a solution to save and organize those video segments with hyperlinks which are relevant to the case for a longer time. All hyperlinks generate a chain of evidence (i.e., the road map in the mental model) to facilitate video analysis.

2.4 Evaluation Index System of Natural Interaction

After the video image investigation system was proposed, another key research question was how to evaluate the naturalness and the efficiency of the system interaction. In this study, work efficiency, mental workload, and emotional experience were considered as three evolution indices.

Work Efficiency. Here, work efficiency refers to an analyst's activities directed toward the accomplishment of video analysis. Time to completion (TTC) and learning time are two common indices of work efficiency. A shorter time to get familiar with an interactive system or to complete a task using this interactive system indicates its better work efficiency. The previous study compared 3 levels of learning time (no learning

time, 15-min and 30-min learning time) and examined the effects of learning time on the work efficiency of three video summarization systems. The authors found shorter TTCs with the longer learning time, suggesting that learning time is an effective indicator of work efficiency [10].

Mental Workload. Mental workload reflects the interaction of mental demands imposed on operators by tasks they attend to [11] or the mental cost of accomplishing the task demands [12]. Subjective ratings, task performance and physiological signals are able to measure mental workload. Among them, EEG signals are sensitive to subtle changes in mental workload. Under the high mental workload condition, alpha band activity is suppressed while theta band activity increases [13–15]. When dealing with a complex or multitasking situation, theta band activity increases at the frontal and parietal areas [16, 17].

Emotional Experience. Emotional experience refers to affective feelings at work. EEG signals are sensitive to the changes in emotional states and frontal alpha asymmetry (FAA) is the most widely used indicators. FAA reflects the differences of alpha band activity between left and right frontal lobes. Alpha asymmetric pattern can be explained by two models: motivational model and valence model. Emotions with approach motivational tendencies are linked to a higher left frontal activity, whereas emotions with withdrawal motivational tendencies are linked to a higher right frontal activity [18, 19]. On the other hand, greater left hemisphere activity (lower alpha power) is associated with positive emotion, whereas greater right hemisphere activity is associated with negative emotion [20, 21].

3 Evaluation of the Proposed Video Image Investigation System

In this section, we conducted a user study to evaluate the naturalness and the efficiency of the system interaction. Only a few comparable features (e.g., sketch-based annotation vs. paper-based annotation) between the proposed interactive system and the traditional method were investigated in the user study.

3.1 Participant

Thirty healthy volunteers took part in the experiment. Due to incomplete EEG data recordings, two participants were excluded from the further analysis, leading to a sample of 28 participants (13 males and 15 females), whose average age was 24 (range = 20–29, standard deviation = 2.3) years old. All had normal or correct-to-normal eye-sight and no previous nervous or psychiatric disorders.

3.2 Task Description

Participants were required to watch thirteen video clips and detect a target suspect from each video clip, which would eventually generate a road map of the target suspect from the first time appeared in the video to the time and place of the case. Because the target

suspect appeared in all video clips and appeared more than one time in some video clips, participants had to watch each video clip from the beginning to the end (jumping forward was not allowed). Normal interactive operations with a video player (e.g., select and open a video clip, play, pause, jump backward) were identical for two groups. Participants in the experimental group used the proposed video image investigation system (Fig. 2a) while those in the control group used traditional screenshots and paper-based annotations (Fig. 2b).

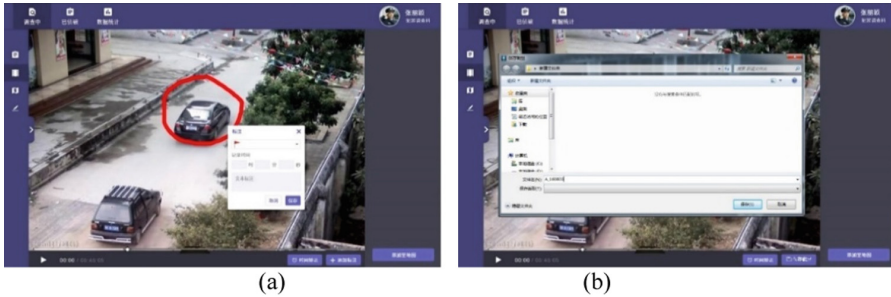


Fig. 2. (a) Interaction with the proposed video image investigation system. When a target appeared in the video (e.g., a black car), participants in the experimental group used the mouse to draw a circle around the target. This operation generated a screenshot automatically and a pop-out window which allowed the editing of annotation information. (b) Traditional screenshots and paper-based annotations. When a target appeared in the video, in contrast, participants in the control group paused the video to save a screenshot and write down all annotation information on the notebook.

3.3 Procedure

Upon arrival, participants completed an informed consent form, followed by a questionnaire regarding their demographic information. After the setup of an EEG cap, they sit in front of a standard desktop computer (3.40 GHz, Intel Core i7 processor) with a 22-in monitor in the lab. The viewing distance to the monitor was approximate 65 cm.

Participants first went through a practice session to get familiar with the interactive investigation system (experimental group) or traditional screenshots and paper-based annotations (control group). After that, participants were instructed to sit quietly and watch a blank screen with their eyes open for 5 min (i.e., baseline condition). Before the formal test, participants were informed about the task and video materials. They were encouraged to finish the task as accurately and quickly as possible. The whole experiment lasted for about 2 h and the participants received ¥100 as reimbursement.

3.4 EEG Acquisition

We used a 64-electrode Neuroscan Cap to collect participants' brain activities when they watched video clips. All electrodes were placed according to the International 10–20 electrode placement standard. The reference electrodes were placed on the left and

right mastoids. The horizontal and vertical EOGs were recorded with electrodes placed 10 mm away from the outer canthi of both eyes and below and above the left eye. Electrical impedances at each electrode site were reduced to less than 5 kOhms. In addition to two reference electrodes, 28 electrodes from the frontal and parietal areas (FP1, FPZ, FP2, AF3, AF4, F5, F3, F1, FZ, F2, F4, F6, FC5, FC3, FC1, FCZ, FC2, FC4, FC6, P7, P5, P3, P1, PZ, P2, P4, P6, P8) were used. The sampling rate was 1024 Hz.

The raw EEG data were digitally filtered with a 0.1–50 Hz bandpass filter. All trials were visually inspected and those trials with excessive peak-to-peak deflections and bursts of electromyography activity were excluded from further analyses. The experimenter looked through a bunch of eye blinks and figured out a threshold that would catch the majority of them for each participant. All the blinks within the threshold were segmented into epochs and rejected in the set that contained more than 1 eye blink or appeared to deviate from the norm. The spatial singular value decomposition was performed to create the ocular artifacts linear derivation file which was used to approximate the topographies for each component to be removed from the EEG raw data. The average of M1 and M2 was used as the reference electrodes. Clean and re-referenced EEG data were transformed into a frequency domain by a short-term Fast Fourier transformation through a 2-s Hanning window. Power values within the theta band (4–8 Hz) and alpha band (8–13 Hz) were averaged for each resting and task conditions for each participant.

4 Results

4.1 Mental Workload

EEG data during the experimental task was subtracted by EEG data at rest, which were normalized into the range [0, 1]. A one-way analysis of variance (ANOVA) was performed with group (2 levels: experimental group vs. control group) as a between-subjects variable. Dependent variables were the theta band power at the frontal lobe and alpha band power at the parietal lobe.

As shown in Fig. 3, significant differences in the theta band power between two groups were found at F4 site ($F(1,26) = 4.774$, $p = 0.038$, $\eta^2 = 0.155$), F5 site ($F(1,26) = 6.126$, $p = 0.020$, $\eta^2 = 0.191$), and FC6 site ($F(1,26) = 5.351$, $p = 0.029$, $\eta^2 = 0.171$). A significant difference between two groups was also found for the average theta band power of the frontal areas ($F(1,26) = 4.251$, $p = 0.049$, $\eta^2 = 0.141$). The frontal theta band power of the experimental group was significantly lower than that of the control group (see Table 1) at these electrode sites. On the other hand, however, no significant difference was obtained for the alpha band power at the frontal or parietal area. These results were consistent with the previous findings of mental workload, indicating that the theta band powers at the frontal areas were sensitive to the rest-task differences in the mental workload. Participants using the proposed interactive system might spend less cognitive efforts/resources to complete the video image investigation task compared to those with the traditional paper-based method.

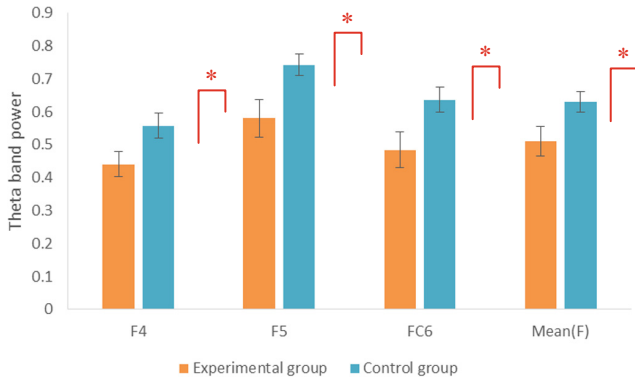


Fig. 3. Comparison of the theta band power at the frontal areas between two groups of participants (error bars indicate ± 1 standard error).

Table 1. Average and standard errors of the theta band power between two group of participants.

| Electrode sites | Experimental group | Control group |
|------------------------------|--------------------|-------------------|
| F4 | 0.441 \pm 0.039 | 0.558 \pm 0.037 |
| F5 | 0.580 \pm 0.057 | 0.742 \pm 0.033 |
| FC6 | 0.484 \pm 0.054 | 0.636 \pm 0.038 |
| Average of the frontal areas | 0.510 \pm 0.045 | 0.631 \pm 0.031 |

4.2 Emotional Experience

A repeated measures analysis of variance (ANOVA) was performed with hemisphere (2 levels: left vs. right hemisphere) as a within-subjects variable and group (2 levels: experimental group vs. control group) as between-subjects variable. The frontal alpha asymmetry (FAA) was dependent variables which were computed with 6 pairs of electrode sites at the frontal areas (FP1/FP2, AF3/AF4, F1/F2, F3/F4, F5/F6, F7/F8). Significant interaction findings were followed-up with the simple effect analysis, in which the differences in the alpha frontal asymmetry between the left and right hemisphere were assessed for each group of participants.

The hemisphere \times group interaction was significant for the pair of F1/F2 ($F(1, 25) = 6.982$, $p = 0.014$, $\eta^2 = 0.206$) and the pair of F3/F4 ($F(1, 25) = 5.519$, $p = 0.027$, $\eta^2 = 0.17$). Simple effect analysis showed significant differences of the FAA between the left and right hemisphere F1/F2 ($p = 0.019$) and F3/F4 ($p = 0.03$) in the experimental group (see Fig. 4). Specifically, the alpha band power at the right hemisphere was significantly larger than that at the left hemisphere in the experimental group, while there was no significant difference in the control group (see Table 2). There was no significant main effect for either hemisphere or group. These results

indicated that participants in the experimental group showed more asymmetric brain activation in the left than right hemisphere, suggesting an increased allocation of the cortical activation. Because the left hemisphere activity correlates with positive affection or approach motivation, participants using the proposed interactive system experienced more positive emotions or approaching motivation during the process of investigation compared to those with the traditional paper-based method.

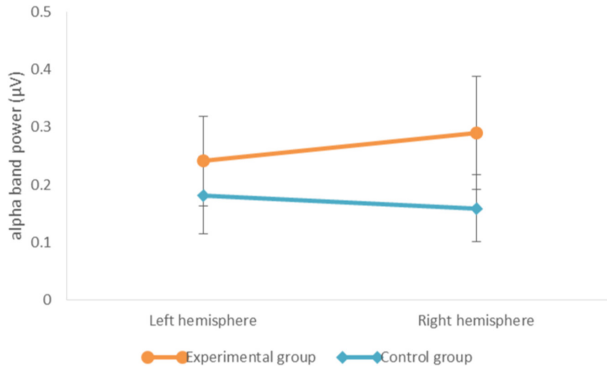


Fig. 4. Comparison of the alpha band power between the left and right frontal lobe between two groups of participants (error bars indicate ± 1 standard error).

Table 2. Average and standard errors of the alpha band power between two group of participants.

| Electrode sites | Experimental group | Control group |
|-----------------|--------------------|---------------|
| F1 | 0.241 ± 0.070 | 0.182 ± 0.042 |
| F2 | 0.290 ± 0.084 | 0.159 ± 0.034 |
| F3 | 0.243 ± 0.067 | 0.211 ± 0.049 |
| F4 | 0.298 ± 0.070 | 0.187 ± 0.037 |

4.3 Performance Data

The time to completion (TTC) was calculated to reflect the task performance. A one-way ANOVA was performed with group as a between-subjects variable. As shown in Fig. 5, there was significant difference in the TTC between the experimental group and control group ($F(1, 28) = 17.391, p = 0.000, \eta^2 = 0.383$). Participants using the proposed interactive system spend less time (mean = 43.033, standard error = 0.972 min) to complete the video image investigation task than those with the traditional paper-based method (mean = 59.3, standard error = 3.778 min).

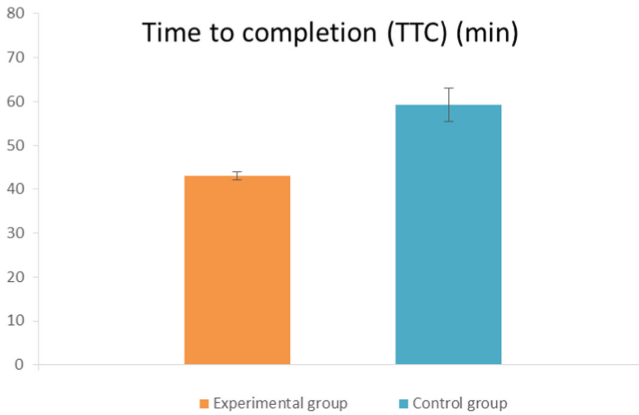


Fig. 5. Comparison of the TTC between two groups of participants (error bars indicate ± 1 standard error).

5 Discussion

This study designed an interactive system based on the mental model of the video image investigation. A user study was conducted to evaluate the naturalness of the newly designed system in terms of work efficiency, mental workload, and emotional experience. Through the analysis of EEG data, we found that the proposed interactive system led to higher work efficiency, lower mental workload, and better emotional or motivational states compared to traditional paper-based method. Specifically, the theta band power increased at the frontal area when high mental demands was imposed on the participant, while no significant difference in the alpha band power was found between two groups. The previous studies showed similar results in the field of traffic monitor and power plant center [22, 23] in which higher theta band power and lower alpha band activity were observed under the high mental workload condition. The relationship between the EEG band power and multiple mental activities (e.g., visual searching, monitoring, attention) may explain the current findings [24].

Analysis of the frontal alpha asymmetry results showed that a significant increase of the alpha activity was observed in the right hemisphere for participants using the interactive system. According to the valence model, the alpha activity in the left hemisphere is positively correlated with positive emotions while the alpha activity in the right hemisphere is positively correlated with negative emotions. The current FAA results indicated that participants with the interactive system experienced better emotional states. Compared to traditional paper-based method, participants using the interactive system felt more positive emotions or they tended to use the interactive system [25]. One possible reason was that the new interactive system can easily add annotations for each target with hyperlinks which help to organize different cases and clues into an integrated part.

In the future work, a field study involving the actual video analysts will offer more powerful results and insights for the improvement of the proposed interactive system.

Professional video analysts with experience and skills in video image investigation may use this interactive system with their own preferences. Moreover, portable EEG devices with real-time recognition of mental workload and emotional states will benefit the evaluation.

Acknowledgments. This work was supported by the National Key Research and Development Plan (2016YFB1001200) and the National Natural Science Foundation of China (31771226). We thank all video analysts who took part in our interviews in the early development of mental model of the video image investigation.

References

1. Slone, D.J.: The influence of mental models and goals on search patterns during web interaction. *J. Am. Soc. Inform. Sci. Technol.* **53**, 1152–1169 (2002)
2. Gentner, D., Stevens, A.L.: *Mental Models*. Lawrence Erlbaum Associates Inc., Hillsdale (1983)
3. Weyman, A., O'Hara, R., Jackson, A.: Investigation into issues of passenger egress in Ladbroke Grove rail disaster. *Appl. Ergon.* **36**, 739–748 (2005)
4. Mack, Z., Sharples, S.: The importance of usability in product choice: a mobile phone case study. *Ergonomics* **52**, 1514–1528 (2009)
5. Larsson, A.F.: Driver usage and understanding of adaptive cruise control. *Appl. Ergon.* **43**, 501–506 (2012)
6. Baxter, G., Besnard, D., Riley, D.: Cognitive mismatches in the cockpit: will they ever be a thing of the past? *Appl. Ergon.* **38**, 417 (2007)
7. Revell, K.M.A., Stanton, N.A.: Case studies of mental models in home heat control: searching for feedback, valve, timer and switch theories. *Appl. Ergon.* **45**, 363–378 (2014)
8. Kriz, S., Hegarty, M.: Top-down and bottom-up influences on learning from animations. *Int. J. Hum.-Comput. Stud.* **65**, 911–930 (2007)
9. Joukes, E., Cornet, R., de Bruijne, M.C., de Keizer, N.F.: Eliciting end-user expectations to guide the implementation process of a new electronic health record: a case study using concept mapping. *Int. J. Med. Inf.* **87**, 111–117 (2016)
10. Liu, Y.J., Ma, C., Zhao, G., Fu, X., Wang, H., Dai, G., et al.: An interactive SpiralTape video summarization. *IEEE Trans. Multimedia* **18**, 1269–1282 (2016)
11. Gopher, D., Donchin, E.: Workload: an examination of the concept. In: Boff, K.R., Kaufman, L., Thomas, P. (eds.) *Handbook of Perception and Human Performance. Cognitive Processes and Performance*, vol. 2, pp. 1–49. Wiley, Oxford (1986)
12. Wickens, C.D.: Multiple resources and performance prediction. *Theor. Issues Ergon. Sci.* **3**, 159–177 (2002)
13. Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., et al.: EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Space Environ. Med.* **78**, B231–B244 (2007)
14. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F.: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* **44**, 58 (2014)
15. Dussault, C., Jouanin, J.C., Philippe, M., Guezennec, C.Y.: EEG and ECG changes during simulator operation reflect mental workload and vigilance. *Aviat. Space Environ. Med.* **76**, 344–351 (2005)

16. Fairclough, S.H., Venables, L.: Prediction of subjective states from psychophysiology: a multivariate approach. *Biol. Psychol.* **71**, 100–110 (2006)
17. Fairclough, S.H., Venables, L., Tattersall, A.: The influence of task demand and learning on the psychophysiological response. *Int. J. Psychophysiol. Off. J. Int. Organ. Psychophysiol.* **56**, 171–184 (2005)
18. Davidson, R.J.: Cerebral asymmetry and emotion: conceptual and methodological conundrums. *Cogn. Emot.* **7**, 115–138 (1993)
19. Davidson, R.J.: Parsing affective space: perspectives from neuropsychology and psychophysiology. *Neuropsychology* **7**, 464–475 (1993)
20. Gotlib, I.H., Ranganath, C., Rosenfeld, J.P.: Frontal EEG alpha asymmetry, depression, and cognitive functioning. *Cogn. Emot.* **12**, 449–478 (1998)
21. Heller, W., Nitscke, J.B.: Regional brain activity in emotion: a framework for understanding cognition in depression. *Cogn. Emot.* **11**, 637–661 (1997)
22. Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A.R., Miyake, S.: Assessment of operators' mental workload using physiological and subjective measures in cement, city traffic and power plant control centers. *Health Promot. Perspect.* **6**, 96–103 (2016)
23. Majid, F., Majid, M., Rashid, H., Ali Reza, S., Shinji, M.: Effects of mental workload on physiological and subjective responses during traffic density monitoring: a field study. *Appl. Ergon.* **52**, 95 (2016)
24. Puma, S., Matton, N., Paubel, P.V., Raufaste, É., Elyagoubi, R.: Using theta and alpha band power to assess cognitive workload in multitasking environments. *Int. J. Psychophysiol. Off. J. Int. Organ. Psychophysiol.* (2017)
25. Briesemeister, B.B., Tamm, S., Heine, A., Jacobs, A.M.: Approach the good, withdraw from the bad—a review on frontal alpha asymmetry measures in applied psychological research. *Psychology* **4**, 247–265 (2013)



An Experiment Study on the Cognitive Schema of Trajectory in Dynamic Visualization

Xiaozhou Zhou¹, Chengqi Xue^{1(✉)}, Congzhe Chen²,
and Haiyan Wang¹

¹ School of Mechanical Engineering, Southeast University,
Nanjing 211189, China

{zxx, ipd_xcq}@seu.edu.cn

² The 60th Research Institute of General Staff Department of P.L.A.,
Nanjing, China

Abstract. In theory, when the novel has a high-matching degree with the cognitive structure, the cognitive process maintains an equilibrium status, the assimilation process of cognitive schema generates at this time. And the new information may expand the existing cognitive structure in amount and strengthen the schema. However, if the novel has a low-matching degree with the existing cognitive structure, the cognitive process appears unbalance accompanied by the accommodation process. In big data era, the visual features of the movement of objects (trajectory) has become an important element in dynamic visualization. In the display space, the movement pattern corresponds to the movement schema in cognitive schema. Based on this, we designed an experiment to verify the theoretical reasoning of cognitive schemas of trajectory in dynamic visualization. The results showed that the cognitive schemas of trajectory in dynamic visualization could build up by iterative learning in a short time. And the cognitive schema had a certain degree of inclusiveness. The difference degree between the novel and the inherent information was the main factor of the effect of the cognitive schemas. But we didn't found the obvious distinction between the assimilation process and the accommodation process of cognitive schemas in the experiment. We also found the different dynamic trajectories associated with the effect of cognitive schemas to a certain degree. This research opened up a new perspective of cognitive schemas for the study of dynamic visualization.

Keywords: Dynamic visualization · Cognitive schemas · Assimilation
Accommodation cognitive load

1 Introduction

1.1 Schema

People have a natural tendency mode, we used to look for recognizable and meaningful mode of the complex pattern. This is an inherent need to find the law of things [1]. When cognitive subject facing with an unknown thing, he/she tends to look for a thinking or action mode in the pre-existing knowledge which matching the new thing. In the field of psychology, it calls the process of mode matching and also the

mechanism of cognitive schema. The schema is similar to what’s usually known as “concept”, it describes the knowledge that organized in some way or in parts, and these frameworks should be filled with concrete contents. When the information input from outside, the brain automatically retrieves available cognitive schemas and generates the assimilation or accommodation process according with the matching degree of novelty and the cognitive schema [2].

Cognitive schema involves the whole process including information acquisition, understanding, memory, reasoning, judgement and problem solving. Cognitive schemas are stored in long-term memory in which including episodic memory and automatic skills, and integrated with the visual system to recognize thousands of glyphs and visual objects. Besides the external environment, the factors which influent the cognitive schema are including the experience, motivation, interests and emotions of the cognitive individual. Therefore, the cognitive schema performs difference between individuals.

As shown in Fig. 1, when the novel has a high-matching degree with the cognitive structure, the cognitive process maintains an equilibrium status, the assimilation effect generates at this time. The new information may expand the existing cognitive structure in amount and strengthen the schema. However, if the novel has a low-matching degree with the existing cognitive structure, the cognitive process appears unbalance accompanied by the accommodation process [3]. Through learning, the nature of the cognitive structure has been changed and the new schema generated to make the cognitive process to a new balance state.

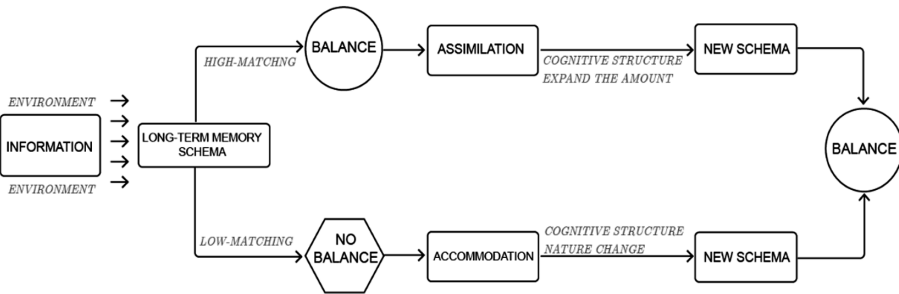


Fig. 1. Cognitive schema mechanism of assimilation and accommodation

To verify the semantic relevance of various semantic schemas, Smith et al. proposed feature comparison model in 1974. They believed the more same features between the concepts, the more connected in semantics. A feature comparison model was employed to further classify features. The essential feature of things was named defining features, and the nonessential but descriptive features was named characteristic feature [4]. In recent years, the studies on the correlation between population characteristics and cognitive schemas started to appear. Jind et al. focused on the changing patterns of psychological cognitive schemas of detached families [5]. Klibert et al. explored the difference cognitive schemas in patients with perihism and generalized anxiety disorder [6]. And Alvidrez et al. found that the schematic representation of ethnic minorities among young college students varied according to their social class [7].

1.2 Cognitive Schema in Dynamic Visualization

However, there were few researches on the cognitive schemas of dynamic visualization. Cognitive schema plays a key role in the process of dynamic visualization. In the era of big data, dynamic representations are now more common due to the increasing scale and dimensions of the information. The dynamic visualization with temporal attribute could guide individual's perception by the features of visual presentation. In the GUI visualization, the data information migrates in visual features according to the temporal dimension. These visual representations are mainly composed of the primitives in spatial structure and visual coding [8, 9]. Thus, the visual features of the movement of objects (trajectory) has become an important element in dynamic visualization. In the display space, trajectory of the nodes can be represented the transition and migration of the data. The movement pattern corresponds to the movement schema in cognitive schema. Combined with the precious empirical structure, it could help us grasp the complex action patterns quickly by mapping the understanding of action schema in low-level to the action information need to learn [10].

Cognitive schema is a black box for us, so we want to understand the mechanism of assimilation and accommodation matching process of cognitive schema to achieve the goal of visual cognition. In theory, assimilation is produced by unilateral access of knowledge which cost less cognitive effort, while accommodation is integrating the relevant schemas to a new knowledge structure which need more cognitive effort for its more complex and varied in structure. Therefore, the experiment attempted to complete a new schema through the pattern implantation of motorial structure.

2 Experiment

2.1 Experimental Design

Specifically, assimilation is a pattern matching process that directly converts the acquired information into the knowledge system of the brain. Then judge the similarity between the acquired knowledge and the knowledge in long-term memory by comparing the definitional and descriptive characteristics of the model. If in high similarity, the novel information could be integrated in the knowledge structure in the form of cognitive schema. Based on this, the design idea of the experiment was to establish a cognitive structure within a short time through the implantable schemas, i.e. experimental materials learning process and then provided the dynamic visual stimulus with feature comparison. The subject would make the decision when compared the similarity between the implantable patterns and the matching stimulus as shown in Fig. 2(a).

Correspondingly, when in low similarity, the subject could not incorporate the novel information into their own knowledge structure in the form of cognitive schema, the subjects should adjust their knowledge structure to a new schema as to conform to the descriptive characteristics of the acquired knowledge. This pattern matching process is accommodation. In the experiment, the changing amplitude was quantitative controlled, and the accommodation effect would be more obvious when differences increased, as shown in Fig. 2(b).

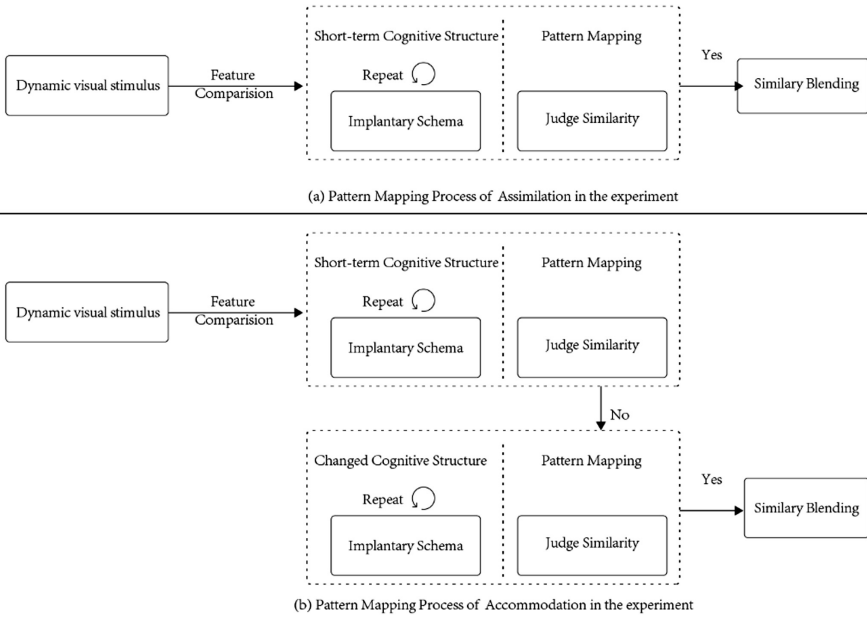


Fig. 2. Pattern mapping process of assimilation and accommodation in the experiments

2.2 Subjects and Materials

The experimental paradigm evolved from the Corsiblocks span task [4], an experimental paradigm that tests visual-spatial cognitive ability. The experimental focused on the cognitive schema of trajectory. As shown in the experimental process in Fig. 3, after displaying the introduction that described the experimental steps and operation

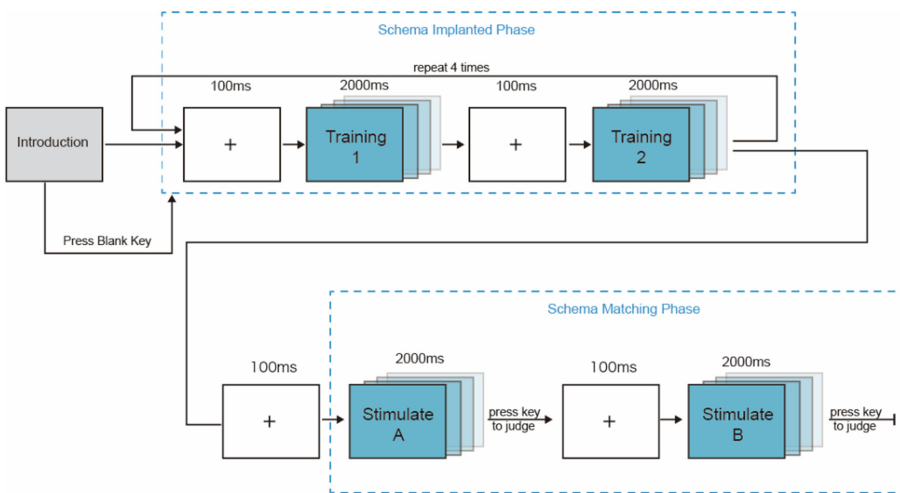
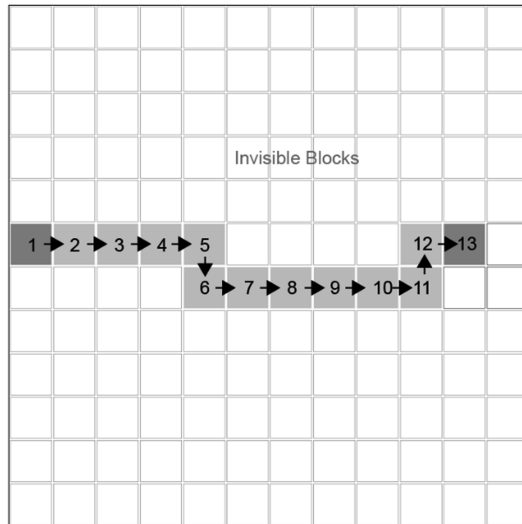
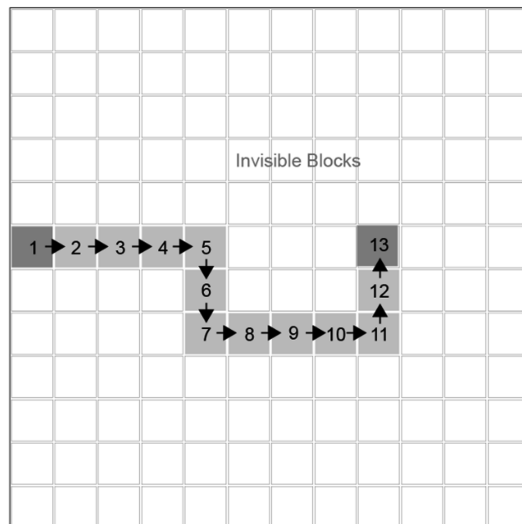


Fig. 3. An experimental trial of the procedure of cognitive schema of trajectory

methods, the screen showed the cognitive schema implantation phase in each trial in which two training materials (Fig. 4) were presented repeatedly. The material 1 corresponded the letter F, while the material 2 corresponded the letter J. After displaying alternations for 5 times, the stimulus material A and B that similar to the material 1 and 2 were displayed in the cognitive schema matching phase. The subjects were asked to judge the attributable of the stimulus materials and press F or J at the first time. And 20



SCHEMA IMPLANTED MATERIAL 1



SCHEMA IMPLANTED MATERIAL 2

Fig. 4. The diagram of cognitive schema trajectory implanted training materials in the experiment

graduate students (5 female) participated in each group. All subjects had normal or corrected-to-normal vision. The Tobii X2-300 compact non-contact eye tracker was employed to collect the eye movements' and performance data.

The Fig. 4 showed a diagram of the experimental materials. Since the experimental material was represented the motion trajectory, the actual experimental material was an animation with 13 frames. The animation showed the movement of a white node on a black screen (in each frame only played one white square), and the Arabic numbers in Fig. 4 indicated the plane position of the white node in each frame. In order to eliminate the effect of the primacy effect and recency effect of the working memory. The

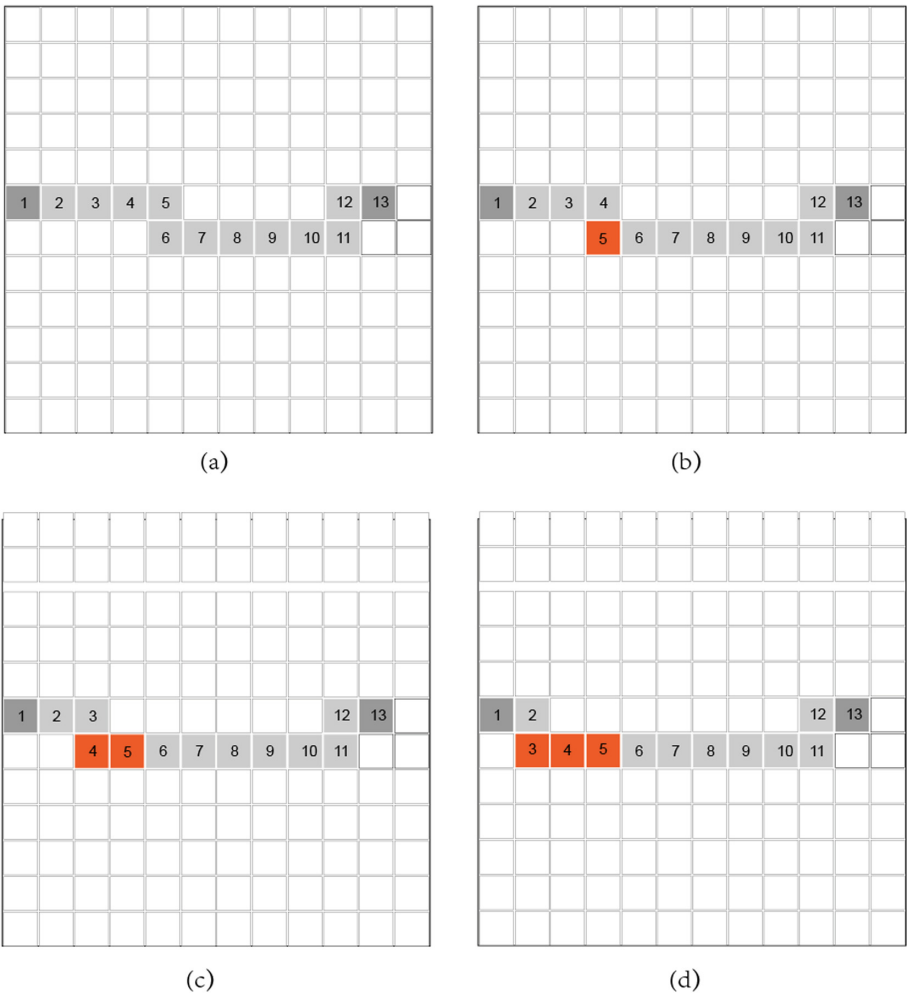


Fig. 5. The quantity change of the trajectory of the dynamic visual stimulus in the experiment. (a) Training material 1 displayed in the schema implanted phase; (b) quantify the degree of difference 1; (c) quantify the degree of difference 2; (d) quantify the degree of difference 3.

initial position and the ending position in the animation were kept uniform between the training material and the similar stimulate material. And the independent variable was the motion trajectory of the node.

In order to investigate the different effects of cognitive schema, the stimulus materials which displayed after the implantation schema was designed in quantitative changes. As shown in Fig. 5, the similar stimulus material of schema implanted material 1 provided three different magnitude changes. In the premise of consistency in the first and last position and the overall path shape (right down right up and right), the number of changed waypoints in the trajectory were 1, 2, and 3, respectively.

3 Results and Discussions

The experimental data including the correct rate, reaction time and eye-tracking were recorded during the experiments. There were two independent variables in the experiments, which were two types of experimental materials (shown in Fig. 4) and three difference degree between the target stimulus and the training materials. The two-factor interaction map with the dependent variable of mean value of reaction time of 20 subjects was shown in Fig. 6. As we seen, there was a cross trend between the two ordinates that indicated the interaction effect between the two factors to a certain degree.

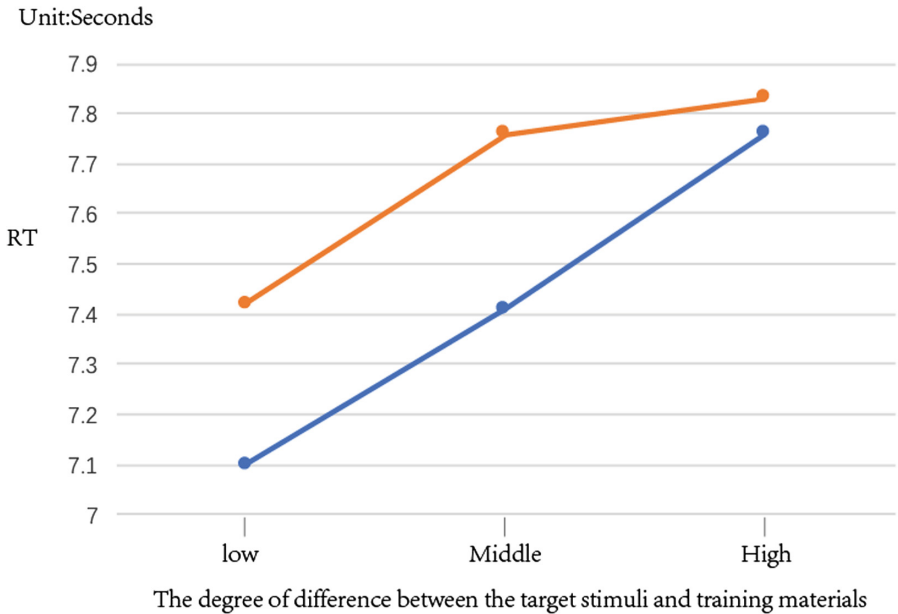


Fig. 6. The interaction map of trajectory cognitive schemata study.

This experiment could reflect the mechanism of the cognitive schema in some extent. The t-test result between two different type of experimental materials $t = -1.702$, $df = 118$, $p = 0.091$, showed main effect was not remarkable. The ANOVA test of the three-difference degree between the target stimulus and the training materials showed the significant effect, $F(2,117) = 3.448$, $p = 0.035 < 0.05$. And when the trajectory difference was in low level (shown as Fig. 5(b)), there was no misjudge in distinguishing the material's attribution. When the difference degree increased, the subjects started error with the rate about 10%. The high correct rate showed that the subjects did generate the cognitive schemas of the dynamic trajectory during the experimental process.

The results of ANOVA test showed that there was a significant difference between three-difference degree in material 1, $F(2,57) = 3.659$, $p = 0.032 < 0.05$, but was no significant difference in material 2, $F(2, 57) = 0.934$, $p = 0.399 > 0.05$. In the training phase of the experiment, the cognitive schemas of dynamic trajectory were generated from the two stimulus materials and the associated letters. And the cognitive schemas determined cognitive performance in the schema mapping phase. The test results of both the interaction of two factors and different significance indicated that even experimental materials effected the change magnitude on the effects of cognitive schemas, but the main factor of performance of cognitive schemas was the similarity between the novel knowledge (schema matching phase) and the intrinsic schema (schema implanted phase).

There were multiple units of similar target materials with same difference degree in the schema matching phase. We compared the eye-tracking data of two units of the same material and difference degree with long interval. Among them, the fixation dwell time and the number of fixations were common eye-tracking indexes which could indicate the immediate cognitive load [11, 12]. The paired t-test results of the same subjects showed that there was no significant difference in the index of fixation dwell time $t = 1.141$, $df = 18$, $p = 0.269$, and also no significant difference in number of fixations $t = 1.161$, $df = 18$, $p = 0.874$. The results indicated that the subjects had no significant changes in cognitive load during the experimental process.

4 Conclusion

Theoretical reasoning and experimental analysis showed that the cognitive schemas of trajectory in dynamic visualization could build up by iterative learning in a short time. And the cognitive schema was not fixed but had a certain degree of inclusiveness. The difference degree between the novel and the inherent information was the main factor of the effect of the cognitive schemas. But we didn't found the obvious distinction between the assimilation process and the accommodation process of cognitive schemas. We also found the different dynamic trajectories associated with the effect of cognitive schemas to a certain degree. This research opened up a new perspective of cognitive schemas for the study of dynamic visualization.

Acknowledgement. This paper is supported by National Natural Science Foundation of China (No. 71471037). Thanks for all the participants involved in the experiments.

References

1. Aiken, R.B.J.A.: Richard Padovan—proportion: science, philosophy, architecture. *Isis* **93**(3), 113–122 (2002)
2. Bartlett, G.C.F.: Remembering. A study in experimental and social psychology. In: *Schlüsselwerke der Kulturwissenschaften* (2012)
3. Morris, R.P.: Method and system for providing a subscription to a tuple based on a schema associated with the tuple, US20090307374 (2009)
4. Rips, L.J., Smith, E.E., Shoben, E.J.: Set-theoretic and network models reconsidered: a comment on Hollan’s “Features and semantic memory”. *Psychol. Rev.* **82**(2), 156–157 (1975)
5. Jind, L., Elklit, A., Christiansen, D.: Cognitive schemata and processing among parents bereaved by infant death. *J. Clin. Psychol. Med. Settings* **17**(4), 366–377 (2010)
6. Klibert, J., Lamis, D.A., Naufel, K., et al.: Associations between perfectionism and generalized anxiety: examining cognitive schemas and gender. *J. Ration.-Emot. Cognit.-Behav. Ther.* **33**(2), 160–178 (2015)
7. Alvidrez, S., Igartua, J., Martinez-Roson, M.: Schematic representations of ethnic minorities in young university students. *Anales De Psicología* **31**(3), 930–940 (2015)
8. Marr, D., Nishihara, H.K.: Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B Biol. Sci.* **200**(1140), 269–294 (1978)
9. Archambault, D., Purchase, H.C.: Can animation support the visualisation of dynamic graphs? *Inf. Sci.* **330**, 495–509 (2015)
10. Berch, D.B., Krikorian, R., Huha, E.M.: The corsi block-tapping task: methodological and theoretical considerations. *Brain Cognit.* **38**(3), 317–338 (1998)
11. Van Orden, K.F., Jung, T.P., Makeig, S.: Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biol. Psychol.* **52**(3), 221–240 (2000)
12. Jacob, R.J.K., Karn, K.S.: Commentary on section 4 – eye tracking in human-computer interaction and usability research: ready to deliver the promises. *Minds Eye* **2**(3), 573–605 (2003)

Cognition in Aviation and Space



Playbook for UAS: UX of Goal-Oriented Planning and Execution

Jack Gale¹(✉), John Karasinski², and Steve Hillenius²

¹ San Jose State University Research Foundation, San Jose, CA 95112, USA
jack.w.gale@nasa.gov

² NASA Ames Research Center, Mountain View, CA 94035, USA

Abstract. We are evaluating Playbook for CASAS (Connected Autonomous Smart Aerospace Systems), a tool designed to aid first responders in disaster relief efforts. We are adapting an existing tool, Playbook, to support a future unmanned aircraft system (UAS) swarm demonstration. Playbook for CASAS will be used to plan, edit, and monitor simulated UAS swarms, and we are interested in evaluating the user experience of this prototype as well as developing recommendations for future UAS interfaces. Allocation of roles and responsibilities between human-automation systems is key to promoting productive cooperation between users and automation. Future interfaces, however, must allow for adaptive management of the swarm not a constant split in human-automation control. Our early research indicates that when a single pilot is controlling swarms of robotic agents, such as UAS or ground rovers, operators require a higher level, goal-based interface with usability at its core. Along with that high-level control, users can leverage sensors within the swarm to be notified when lower level actions must be taken by the pilot. First responders working in disaster relief efforts require a high level of situational awareness (SA) and precise control at key moments within a mission. This balance in operator workload paired with SA can lead to improved safety and mission outcomes. Our research below outlines leverage points as well as the balance between human involvement and autonomy in UAS interfaces.

Keywords: Swarms · UAV · UAS · Autonomy · Engineering psychology
Cognitive ergonomics · User experience · Human centered design · HCI

1 Introduction

1.1 Future UAS

The field of unmanned aircraft systems (UAS) have seen tremendous growth over the past few decades. Recent advances in hardware have reduced the size and cost of UAS, spurring greater research interest in both government and private industry. The resulting increased availability in this technology has led to a need to provide untrained operators (i.e., operators without explicit pilot training) the ability to monitor and command both individual and swarms of UAS. As the number of UAS in a swarm increases, however, it is important to provide an operator with a user interface that maximizes gains in

situational awareness (SA) while simultaneously minimizing the increases in workload associated with managing the additional vehicles. We have designed and evaluated a novel interface for UAS swarm management which requires minimal-to-no training for use by first responders for disaster relief. Using a human-centered design usability study, we have evaluated this interface with first responders, analyzed their feedback, and synthesized recommendations for similar swarm interfaces.

1.2 Situational Awareness

UAS have been embraced by emergency and disaster relief communities for their use in situations where it is either impractical, impossible, or extremely dangerous to deploy first responders. These vehicles can monitor emergency areas, deliver supplies, and search for missing persons, often faster than would be possible with humans and with no risk to their operator. Perhaps more importantly, these vehicles can be used in tandem with first responders already on the ground, providing greater situational awareness and safety in emergency situations. This increase in situational awareness allows first responders to more effectively complete their jobs.

Endsley defines situational awareness as “the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” [1]. While UAS can provide important SA to first responders on the ground, this should not be a purely unidirectional information flow. The information first responders glean on the ground can be used to redefine the goals of the assisting aircraft. Some tasks, such as modifying the flight plans of UAS, however, requires that first responders also take on the role of air traffic controllers (ATCs). Endsley specifically calls out the “taxing” role of air traffic controllers when she notes that controllers “must maintain up-to-date assessments of the rapidly changing locations of aircraft...and their projected locations relative to each other” [1]. This requirement cannot come without some increase in workload, but a well-designed interface should minimize this impact. The ability to increase situational awareness through the use of live video and other sensor feeds can provide ground workers the knowledge they need to remain safe in highly dynamic scenarios.

1.3 Goal-Oriented Planning

In an ideal scenario, first responders, using the contextual information they’ve gained on the ground, should be able to control the UAS by pointing them to regions of interest and commanding them to accomplish goals. First responders can use goal-oriented planning in order to minimize the impact of this additional role as ATC. Goal oriented planning can be defined as “the process of breaking down complicated goals into simpler, more manageable sub-goals” [2]. As an example, a first responder may have the goal to search a large fire zone for survivors. This goal can be divided into sub-goals by creating a trajectory of several waypoints which cover the entire fire zone. The goal is then completed when the swarm has searched each of the waypoints. In this scenario, the immediate benefit to the first responder is that they need only assign the system to

“search the fire zone” instead of determining and then assigning a complete set of waypoints for the swarm to search.

Goal oriented planning also simplifies the planning process when a first responder has many different goals that need to be completed. As an example, a first responder needs to plan out three goals: “search the fire zone”, “deliver a package”, and “monitor the waypoint”. While each of these goals may contain many sub-goals, the first responder only needs to consider the functional aspects of the goals, such as:

- How long will the goal take to complete?
- How many UAS are required to complete the goal? and
- Which goal has the highest priority?

The first responder can use the answers to these questions to create their plan, without needing to worry about which aircraft will fly to which waypoint at any given time. This allows the first responder to more effectively manage the swarm to respond to current priorities, without needing to concern themselves with the finer details of ATC. This goal level planning has been used by several researchers, with varying numbers of UAS, to result in reduced simulated mission time [3, 4].

2 Interface

2.1 Playbook

Our prototype interface is adapted from the Playbook tool. Playbook is a mobile web-based schedule planning tool currently used in operation by NASA spaceflight analog missions as well as a platform to research crew autonomy. Originally developed as a plan viewer, it was extended significantly in 2014 to become a tool used to enable crew autonomy. These features and capabilities are necessary as we prepare in the spaceflight domain for future deep space missions, where traditional mission control tasks will need to shift to the crew. Shifting these tasks is challenging, as the current roles of mission control are handled by teams of people (100+ in total), where the number of crew onboard may be on the order of 4–6 [5, 6].

The use of the horizontal timeline differs significantly from most consumer products where a vertical timeline is used, primarily because of user familiarity with historical mission plans. Apollo mission plans were produced using a vertical timeline layout, but Shuttle and International Space Station mission plans used a horizontal layout [7, 8]. As our main target users for the Playbook interface and earlier tools were flight controllers and astronauts who were already accustomed to horizontal mission plan timeline layouts, this was used as a basis for retaining their familiarity. Over the years, however, Playbook team has looked at how the same information would be displayed in a vertical layout. Although we have not done a formal study on the differences, our initial impressions, as we tried to represent the high-density space station or Mars rover plans, is that the readability and flexibility break down in such a scenario [9]. The vertical time layouts tended to rely on vertical text which is difficult to read. When data plots such as power usage or altitude, are displayed on a vertical timeline they become difficult to read and to compare with the timeline. The other main consideration in favor of a horizontal

timeline is the heavy use of hierarchy in a mission plan. Each element on the timeline may contain several additional elements (similar to Gantt style charts) that may represent individual activities inside a larger group, or smaller activities that are contained in a larger activity such as on a rover or orbiter mission plan. Three or four levels of hierarchy are not uncommon in these dense mission plans. Since the timeline itself is horizontal, the hierarchy can be displayed vertically, allowing for a natural mapping: higher abstraction are higher vertically compared to lower-level activities. These relationships also map to users familiar with Gantt charts.

2.2 Adapting to UAS

The advantages of providing astronauts with mobile web-based planning tools have benefits and applications that translate over to the UAS space. To support crew autonomy, Playbook was designed with an emphasis to provide lightweight plan editing that allows astronauts to quickly and easily manipulate their mission plan while not taking up time from execution of their spaceflight tasks for the day. Key design goals to enable this were the emphasis on mobile interfaces that allow astronauts to perform mission plan changes without being at a designated computer or console. Emphasis on “walk up and use” usability, and the ability for multiple astronauts to collaboratively work on their mission plan simultaneously was key [5]. These design goals map well to the UAS domain for our use case of first responders arriving at a disaster area. Mobility is key in this UAS context, as our target users are in the field and are not situated at a dedicated console. The target users are not expert pilots and are under similar operational pressures of astronauts, so walk up and use plays a large role in the design and ability to quickly make plan changes. Collaboration allows multiple users to aid in planning, rather than relying on a single user to be designated or otherwise able to accomplish the task at a given time. In addition to the mission planning design elements, the application to the UAS domain required us to integrate time domain planning (used in the Playbook spaceflight mission planning tool) with geospatial planning, which is traditionally seen in many UAS style interfaces. This is represented in the UI by the timeline portion of the interface side by side with the geospatial interface. Depending on what task the user is working on, the timeline or geospatial portion of the interface will grow to maximize the screen real estate, while the non-active portion of the interface will shrink, but not disappear, to allow the user to see the relationship of their planning actions in both the time and geospatial planning domains.

2.3 Interface Elements

Our prototype interface included three main interactive areas which allow users to execute tasks (see Fig. 1). The top of our interface includes a timeline which is very similar to Playbook, but, replaces each crew member row with a single UAS. Each UAS has its own row, followed by a timeline of that UAS future goals and waypoints. Each hour is labeled with a vertical line so users have a sense of when a waypoint or goal will be met or completed.

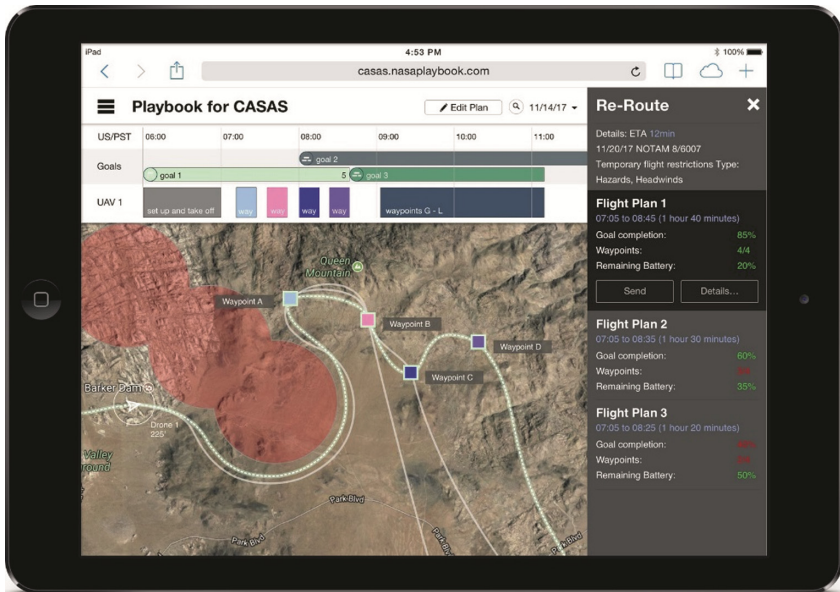


Fig. 1. Playbook for CASAS prototype interface showing the timeline (Top), geospatial view (Center), and stream view (Right). This is showing alternate flight trajectory options, waypoints, and temporary flight restriction (TFR) volumes.

The second area is an overhead or geospatial view, which presents the most radical change to the existing Playbook interface, and much of the focus of our usability studies. This area allows a user to zoom in and out of an area to keep an eye on all UAS, as well as to view trajectories, waypoints, and flight restriction volumes. Users can tap, pinch to zoom, and pan using traditional gestures, although this functionality was not included in our initial prototypes. By leveraging interactions from other widely used products, we can reduce onboarding and training time, especially with the user group of untrained operators.

The last area is the “stream view”. In Playbook, the stream view allows users to glean more detailed information, such as temporal information, constraints, and descriptions of each activity. In this prototype, the stream view is used similarly for waypoints. Some additional functionality was also implemented to this area, which makes it, along with the overhead view, the two primary areas of interaction. The stream view appears on the right side of the screen, and slides in and out as needed. It was designed to be visually distinct from the rest of the interface in order to prompt the user to make decisions at key moments within the mission.

2.4 Early Iterations

Early paper prototypes were employed to better understand key elements, button sizing, placement, and to provide a starting point for discussion with our target users. There are often difficulties early in the design process when it’s unclear what functionality should

be included or removed after initial exploratory research. A paper prototype, which in this case was simply a cardboard cutout of an iPad with paper taped on top, is useful in the early stages of design (see Fig. 2). Paper prototypes allow quick understanding of the hierarchy of elements and the need for certain actions to be added or removed from the interface. One example of this occurred early in this research with the first version of this prototype. We were interested in displaying UAS prognostics to the user and had the idea to show these metrics for each UAS in a card format. This technique is commonly used in interfaces for single UAS pilots. Once we built this into a paper prototype, however, we quickly realized that allocating prognostics space for each drone was unnecessary and consumed too much space on the iPad. The paper prototype allowed us to recognize those areas that needed to be autonomous, and what information might be unnecessary when guiding multiple UAS. This prognostics card idea works well for individual aircraft management; however, that information is much less relevant to a first responder in the field when guiding a swarm.



Fig. 2. Early paper prototype of a swarm interface. This version has more controls which we later removed like velocity and drone allocation sliders.

3 Methodology

3.1 Research Strategy

For this study, our team built an interactive prototype that was presented to subjects with and without training in UAS management. We then describe a “search and recover” mission using a swarm of aircraft that they must control. We asked each participant to complete tasks which include starting the mission, selecting waypoints, and selecting alternate flight trajectories developed by an auto resolver (AR) algorithm. An auto-resolver is defined as a “trajectory-based conflict resolution algorithm that provides efficient flight path changes to solve medium term conflicts” [10]. The auto-resolver used for our upcoming flight demonstration will produce multiple flight trajectories autonomously and use our interface to present these trajectory options to a user. Along with evaluating trajectory selections, we also sought to better understand the role of first responders in the field. To accomplish this, we completed several rounds of early flight tests with three UAS at the NUARC (NASA UAS Autonomy Research Complex), (see Fig. 3).



Fig. 3. Initial testing at NASA Ames Research Center in Mountain View, CA. Three drones completing routes planned through the Playbook interface.

To assist us in our research, we recruited seven subjects made up from members of the Ames Disaster Assistance Rescue Team (DART) and UAS researchers from the Human Systems Integration Division at Ames. Our strategy was to evaluate our prototype interface and better understand our users through the analysis of qualitative data from our usability studies. After testing with subjects, we separated data into three categories: our observations, subjective thoughts from the users, and direct quotes. At the

end of this first round of testing, we developed key insights and recommendations that both inform and evaluate the design of our prototype as well as future unmanned aircraft systems. Our team's primary method of research is a standard human-centered design iterative evaluation loop that allows for rapid prototyping of interfaces, thereby allowing for multiple rounds of usability testing.

3.2 Testing Protocol

Future UAS studies tailored for users with little-to-no experience in UAS management require building a usability protocol with onboard training. Clear usability protocols are essential to familiarize users with UAS controls and complex disaster relief scenarios. Our usability protocol consisted of defined roles for the subjects, an outline of timing and schedule for testing, an introduction, scenario, tasks for the user, completion criteria, and follow up questions. This structured protocol was created to ensure that the data resulting from these studies was as reliable as possible once we began synthesis of our data. We asked subjects to complete a series of tasks that were framed as messages from a ground control station. During their execution of the task, the subject was asked to think-aloud and try to complete the task on our iPad prototype. A think-aloud study asks the user to verbalize each action they take as well as their thinking while completing tasks [11]. This method allows testers to more easily identify and record pain points or moments of confusion, as well as spot differences in their actions versus their intent. The rationale behind conducting a think-aloud usability study is not only to assist testers in highlighting moments of confusion but also times at which the interface felt most clear to the user. Given the increase in use and availability of autonomous systems, NASA has stated an interest in building requirements for testing human-autonomy interfaces and creating verifiable testing protocols [12].

3.3 User Group and Goals

As noted above, this research was conducted with seven volunteer subjects which consisted of Ames Disaster Assistance Rescue Team (DART) members and UAS researchers from the Human Systems Integration Division at Ames. DART is a federal emergency response and recovery team that can be deployed in support of disaster relief efforts. Previous deployments include the Loma Prieta Earthquake, Oakland Firestorm, Oklahoma City Bombing, and the World Trade Center terrorist attack [13]. Many of our subjects are also members of the CA Task Force 3, a FEMA Urban Search and Rescue Task Force which has separate deployments and training exercises also related to disaster relief. The goal of this research was to gain insight into the human's role in an autonomous UAS traffic management (UTM) systems. We are investigating a network of ground control systems which together produce flight path trajectories that ultimately display to users on an iPad. Future interfaces must allow for sufficient human agency, but a balance between human involvement and automation is at the core of future UAS operations. The innovation in our human-autonomy research was to determine those tasks that can be offloaded to the air traffic management systems and those that must be

pushed to the user on the ground. To do this, we must better understand the role of first responders and the needs of this group on the ground.

4 Results

4.1 Introduction

After the first round of testing, we developed a list of insights, recommendations, and behaviors that will inform both our prototype and other unmanned aircraft interfaces. These recommendations and insights highlight user behavior when guiding swarms and underline key leverage points as UAS becomes more ubiquitous. This is by no means an exhaustive study, but is a first step in understanding this complex system within unique mission constraints. It should be noted that these recommendations are pointed towards first responders and firefighters working in disaster areas using unmanned aircraft technology and specifically leveraging swarms.

4.2 Recommendations

Any interface that is designed for walk up and use should provide clear language, and removing air traffic control jargon like NOTAMs is required when designing usable interfaces. NOTAMs are notices to airmen concerning conditions or airspace or hazards usually resulting in a temporary flight restriction and reroute. Our target group has minimal ATC and flight knowledge, so removing this terminology and replacing it with clear language is key to reduce confusion. Another point of confusion with our prototype was that guiding many UAS at once increased the amount of information the user must hold in their head. Clear labels on zones, trajectories, and flight plan options reduce cognitive loads and are essential to assist the user when guiding multiple UAS. Tasks within the geospatial view were also found to be difficult for users because of a lack of clear labels associated with goals and trajectory options. When a user has to hold multiple items in their working memory, like goal completion percentage, waypoints met, and ending battery life for three trajectories, an interface needs to be clear and easily digestible at a glance.

The ultimate goal of the UAS swarm is to increase a user's situational awareness with information gathered by sensors on the UAS. With this in mind, we asked the subjects which sensors would be the most useful during a disaster relief scenario. While subjects responded with a variety of different sensors, they all felt that a live video feed was the most important. Subjects also suggested many other sensors such as heat and radar. Subjects thought heat would be most useful for locating survivors in damaged buildings, and that radar would be useful in gaging heights of those buildings and terrain. While small UAS payloads are extremely limited, we feel that this unique mission environment requires the use of multiple sensors to assist first responders. Providing this sensor data gives first responders more information and tools, enabling greater agency. Users desire agency, and allowing them to control sensors and make real-time decisions on the ground provides freedom to an otherwise fully autonomous system. During our usability study, one participant said, "Pilots don't like being told what to do," and this

thought is an important debate between mission planners and operators. Any human-autonomy partnership suggests minimal human involvement from mission planners, however, operators in the field will commonly request more freedom. Current technology is unable to handle every edge case and critical moment, and it's at these times where the human can take over. Real-time sensor data is also key in enabling situational awareness (SA), especially when evaluating a disaster area. Users first need an overall, 360° view, then a more detailed view of objectives and points of interest. One subject said, "Scene assessment is the first step, start out high, then focus in", this global SA is critical in high-intensity disaster relief environments.

4.3 Behavior

During testing, we noticed behavior from subjects that informed our thinking surrounding UAS interfaces as well as the choice architecture presented in this interface. We found that subjects hesitated before sending instructions to the swarm and ground control stations (GCS). Go/No-Go indicators were recommended by multiple subjects, and could reduce hesitation when sending alternate trajectories or waypoints to GCS. One subject who had experience operating UAS said, when asked about his hesitation when sending trajectories, "It's a military thing, double check everything". Both his behavior and his thinking at that moment led us to believe that greater affordances were needed to assist users in making these decisions.

During our simulated test, subjects were presented with a temporary flight restriction that required them to change the route of a UAS. After selecting their priority waypoints, the prototype suggested three alternate flight plans for the subjects to choose from. Subjects were presented with a geospatial view of the flight plans, as well as data presenting how long each flight plan would take, what percentage of their goal would be met, and the resulting battery life available after the flight plan was completed. Some users placed importance on hitting waypoints or high percentage current goal completion. Others placed importance on saving battery life for future or upcoming goals. This led us to believe that a clear pre-mission briefing was essential for first responders to prioritize mission objectives. In a real disaster scenario, first responders would have clear objectives, and while we did not instruct them to prioritize time over battery life or mission completion, this prototype allowed them to choose flight paths based off their own experience. It also indicated that the interface was successful in allowing for immediate usability, despite the brief pre-flight introduction.

4.4 Patterns

Several patterns within our data emerged after our usability testing was complete. In particular, when subjects were presented with a temporary flight restriction that required them to select high priority waypoints, users felt they didn't have enough information to make an informed decision. Many subjects requested the time to the next waypoint as well as the time from waypoint to waypoint. We also found that users preferred using both the geospatial and stream view to make these waypoints and trajectory selections. While the timeline view of Playbook presented users with this temporal waypoint

information, this was not sufficiently clear, and future prototypes need to address this deficiency.

Along with temporal waypoint information, subjects also asked for weather and wind data to assist them in making future trajectory decisions, as rotorcraft operated by batteries are especially susceptible to weather and wind. If a UAS is flying into a headwind, that information is crucial to a first responder in the field. Lastly, feedback from our usability studies showed an aversion to standard barometric altitude based on sea level. Users felt that absolute altitude (Above Ground Level, or “AGL”) was preferred. One of pilots’ biggest concerns when dealing with low flying aircraft is running straight into objects without knowing it. AGL has a stronger relationship to topography and ground, which is preferred, especially in the case of mountainous areas where wildfires commonly take place.

4.5 Future Steps and Discussion

We also identified some recommendations that are important, but beyond the technical scope of our upcoming flight demonstrations. Through both feedback from our subjects and observations from testing, we identified that first responders in some cases need to create their own trajectories and requested the ability to drag waypoints. This is preferred when negotiating temporary flight restriction (TFR) volumes as the areas can change from moment to moment - especially when these are caused due to the wildfire itself. Auto-resolver algorithms may not be able to send alternate trajectories fast enough and will not have the same information as a first responder on the ground. For this reason, at critical points such as this, we can offload some waypoint and trajectory planning to the first responder when necessary. Users also need more pre-flight information regarding objectives to make these decisions. This could be a preflight briefing as outlined previously, however, future UAS interfaces should be able relay this briefing back to the user when requested during a mission. After our first round of testing, we realized an additional application for swarm technology was to use UAS to create communication network arrays. One subject during our usability testing described the difficult terrain common in wildfire or earthquake scenarios, which can disrupt standard communication. UAS swarms can provide reliable networks for communication when outfitted with appropriate sensors and antenna arrays [14]. However, due to the cost and weight of this equipment, we have not yet been able to incorporate this technology in future flight tests.

5 Conclusion

5.1 Summary of Results

This research is by no means exhaustive, and is only our first round of testing in preparation for our upcoming flight test. We feel that the insights gathered in this initial research provide cues for future interfaces for both first responders and usable swarm interfaces. Allocation of roles and responsibilities between human-automation systems is key to promoting productive cooperation between these two agents. When guiding a

swarm, low-level control should be automated, but, UAS should alert pilots and hand over control at key points within a mission. First responders and firefighters working in disaster areas using unmanned aircraft technology and swarm interfaces need to have a balance between autonomy and workload.

Real-time sensor data is also useful in building situational awareness and notifying users when they should take over control of a single aircraft. Users first need an overall, 360° view, then a more detailed view of objectives and points of interest. First responders need to be able to form a mental model of disaster areas to make informed decisions. These sensors can indicate moments within the mission to notify the user when they should take control of single aircraft. Transferring control over a single aircraft to the operator is useful at certain key moments in a mission, however, this manual process can only be performed on a limited number of aircraft at one time [15]. Once decisions need to be made, Go/No-Go indicators could reduce hesitation and confusion in these high-intensity environments. One other method for reducing confusion and cognitive load is to ensure that these interfaces are removing system-oriented terminology, as first responders often do not have ATC or flight knowledge. Removing technical terminology is key to reduce confusion and promote risk reduction. Finally, breaking complex missions into clear goals frame UAS tasks and mission objectives. As the number of UAS and tasks increase within a mission, clear goals provide clarity for first responders.

One of NASA's human research goals is to develop design guidelines for effective human-automation-robotic systems, specifically distributed swarm systems [16]. The use cases for this type of technology include transportation, construction, relief efforts, military, and space applications. Reducing operator workload in these disaster relief environments can lead to improved safety and mission outcomes. The interface presented here is focused on a future flight demonstration at NASA Ames, however, these insights and recommendations can be translated to other swarm applications outside of the use in disaster relief.

When one user is controlling multiple actors in a system, they are no longer concerned with the low-level actions of each aircraft. Goal-oriented planning and higher-level interactions are more useful [17]. This style of goal-oriented planning can reduce cognitive load for the user as they will no longer have to focus on many concurrent tasks being executed by the swarm. Controlling swarms of robotic agents, such as UAS or ground rovers for space applications, requires a higher level, goal-based interface with usability at its core.

Acknowledgments. The authors would like to thank Jimin Zheng, Matthew Chan, and Richard Joyce for support and development work during the various stages of this project. The authors would also like to thank the subject testing participants, Ames DART, as well as Eric Mueller for his leadership on the Connected Smart Aerospace Systems project (CASAS). This work was performed under a US Govt. Contract in the Human-Systems Integration Division at NASA Ames Research Center.

References

1. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. *Hum. Fact.* **37**(1), 32–64 (1995)
2. Barber, K.S., Han, D.C.: Multi-agent planning under dynamic adaptive autonomy. In: *International Conference on Systems Man and Cybernetics*. Institute of Electrical Engineers Inc. (IEEE) (1998)
3. Jung, D., Ratti, J., Tsiotras, P.: Real-time implementation and validation of a new hierarchical path planning scheme of UAVs via hardware-in-the-loop simulation. *J. Intell. Rob. Syst.* **54**, 163 (2009)
4. Rasche, C., et al.: Combining autonomous exploration, goal-oriented coordination and task allocation in multi-UAV scenarios. In: *2010 Sixth International Conference on Autonomic and Autonomous Systems (ICAS)*. IEEE (2010)
5. Hillenius, S.: Designing interfaces for astronaut autonomy in space. In: *CanUX 2015*, Ottawa, Canada (2015). (Invited Speech)
6. Hillenius, S., Marquez, J., Deliz, I., Kanefsky, B., Korth, D., Healy, M., Gibson, S., Zheng, J.: Designing and building a crew-centric mobile scheduling and planning tool for exploring crew autonomy concepts onboard the international space station. In: *ISS R&D Conference 2016*, San Diego, CA (2016). (Speech)
7. Space Shuttle Program Flight Data File. https://www.nasa.gov/centers/johnson/pdf/567071main_FLT_PLN_135_F.pdf. Accessed 9 Feb 2018
8. Apollo 11 Flight Plan. https://www.hq.nasa.gov/alsj/a11/a11ftpln_final_reformat.pdf. Accessed 9 Feb 2018
9. Hashemi, S., Hillenius, S.: @NASA: the user experience of a space station. In: *SXSW Interactive 2013*, Austin, TX (2013)
10. Erzberger, H.: Automated conflict resolution for air traffic control. In: *Proceedings of 25th International Congress of the Aeronautical Sciences (ICAS)*, Germany (2006)
11. Ericsson, K.A., Simon, A.H.: *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge (1993)
12. NASA: HCI-05.: We need verifiable requirements that specify standard measurement techniques and metrics for evaluating the quality of user interfaces (2017). <https://humanresearchroadmap.nasa.gov/gaps/gap.aspx?i=329>. Accessed 7 Feb 2018
13. DART Homepage. <http://dart.arc.nasa.gov>. Accessed 9 Feb 2018
14. Palat, R.C., Annamalau, A., Reed, J.R.: Cooperative relaying for ad-hoc ground networks using swarm UAVs. In: *Military Communications Conference, MILCOM 2005*. IEEE, Atlantic City (2005)
15. Prevot, T., Homola, J., Mercer, J.: Human-in-the-loop evaluation of ground-based automated separation assurance for NextGen. In: *26th Congress of International Council of the Aeronautical Sciences (ICAS)*, Anchorage, Alaska, USA (2008)
16. NASA: HARI-02 We need to develop design guidelines for effective human-automation-robotic systems (2017). <https://humanresearchroadmap.nasa.gov/Gaps/gap.aspx?i=334>. Accessed 9 Feb 2018
17. Lieberman, H., Espinosa, J.: A goal-oriented interface to consumer electronics using planning and commonsense reasoning. In: *Proceedings of 11th International Conference on Intelligent User Interfaces - IUI* (2007)



Augmented Reality in a Remote Tower Environment Based on VS/IR Fusion and Optical Tracking

Maria Hagl¹(✉), Maik Friedrich¹, Anne Papenfuss¹, Norbert Scherer-Negenborn²,
Jörn Jakobi¹, Tim Rambau¹, and Markus Schmidt¹

¹ DLR German Aerospace Center, Lilienthalplatz 7, Brunswick, Germany
Maria.Hagl@dlr.de

² Fraunhofer IOSB, Gutleuthausstr. 1, Ettlingen, Germany

Abstract. Over the past years, several augmented reality features have been developed to make Remote Tower Operations more cost-efficient and user-friendly. In the context of a national research project (The paper reports results gained in the project “INVIDEON” (FKZ 20 V1505A) sponsored by the Luftfahrtforschungsprogramm (LuFo) of the Federal Ministry of Transport and Digital Infrastructure Germany.), augmented reality based on visual spectrum (VS) and infrared (IR) fusion and as well as on optical tracking is a study objective. Having both VS and IR information available at any time is expected to enable more efficient air traffic control, even at restricted visibility conditions. Integrating VS and IR in one video panorama should also decrease head-down times and therefore increase situation awareness and reduce workload. The integration of two different sensors will be realized by overlaying VS/IR combined with adapted input devices and optical tracking methods. Developing a good concept for the integration of VS/IR and testing it in an exploratory manner can only be achieved with the help of system experts and rapid prototyping methods in simulation environments. During three workshops, human factors specialists, project partners and seven air traffic controllers worked out a prototype that was gradually improved over time and helped to generate a first concept. Firstly, this paper addresses the challenges of VS/IR fusion, manual PTZ following (as a precursor for optical tracking) and adapted input devices. Secondly, it presents the construction process of a prototype in an explorative manner, based on a user-centered approach and implemented in a simulation environment. Finally, it summarizes and presents the results from the workshops and throughout the construction process.

Keywords: Augmented reality · Remote Tower · VS/IR fusion · PTZ control

1 Introduction

Controlling air traffic from anywhere than from a local tower is the core of Remote Tower Operations (RTO). Thanks to optical visual representation of the out-of-the-window view in a digital video panorama, one or more aerodromes can be controlled remotely from a Remote Tower Center (RTC). Originally conceived at the German Aerospace Center (DLR) Brunswick [1] to provide air traffic control (ATC) at a better

cost-efficiency ratio, the idea of remotely controlled air traffic went viral and was firstly operationalized in 2015 [2]. Especially for regional airports, struggling with financial issues, RTO represents an efficient solution. Next to economic benefits, RTO could even outperform conventional tower control, thanks to assistance systems that could support air traffic controller officers (ATCOs) in the future. In this context, the German research project “INVIDEON” investigates how to assist ATCOs in RTO. More specifically, it concentrates on augmented reality based on the fusion of VS and IR images on the digital video panorama (output). A second research question investigates how to support this new work environment with adapted input devices (e.g. control of fusion level, extended use of pan-tilt-zoom function). In this paper, we will at first give a theoretic background on augmented reality in ATC, on VS and IR advantages as well as on input devices used in RTO environments. Later on, methods and contents for the three INVIDEON-workshops are described. Finally, we present current results and give a general prospective for further research.

2 Augmented Reality in Air Traffic Control

Without any system providing augmented reality, ATCOs would only perceive what they could perceive relying on their biological senses. By contrast, augmented reality allows their users to perceive more stimuli than they would actually do through supplementary information (e.g. visual cues) about his or her environment. For ATCOs, who rely especially on their visual faculties to perform their daily tasks at work [3], augmented reality has the potential to provide valuable assistance. Past research has already developed new concepts for augmented reality in conventional tower control. Through head-mounted displays [4–9] or holographic screens [10] ATCOs can be provided by supplementary information they would not see through the out-of-the-window view. Concerning RTO, implementing augmented reality seems to be even easier than in a conventional tower environment. Given that RTO are already based on the visual presentation of an aerodrome in a digital video panorama, features like aircraft detection/identification- or aerodrome information, like weather, wind or stop bars could directly be integrated in the video panorama [11]. Thus, latency between the occurrence stimulation and the display response, that is likely to appear with optical see through displays, can be reduced [12]. With “Head-up Only”, Papenfuss and Friedrich [13] designed a concept aiming for the increase of visual attention through additional information in the video panorama (e.g. approach radar, pan-tilt-zoom camera (PTZ), electronic flight strips, coupled radio frequency, weather data). Due to decreased head-down-times in such a working environment, ATCOs are estimated to work more efficiently since the changing accommodation of the eyes ceases. The anticipated benefits become even more pertinent, when visual information is deteriorated or even inaccessible, due to bad weather conditions or at nocturnal times. At this point, INVIDEON kicked-off.

3 INVIDEON

In the context of further development of RTO, INVIDEON aims for improvement on the design of the video panorama through augmented reality, using optical sensors only. Currently, the standard set-up for RTO is a VS video panorama that represents the out-of-the-window view from a conventional tower. Furthermore, ATCOs use a PTZ as a replacement for conventional binoculars to magnify distant objects of interest. As extension to the standard set-up, some RTCs present IR information on extra screens to get supplementary information when visual conditions are altered. To seize the advantages of having both visual information materials at disposition when needed, the next paragraphs aim to point out the characteristics of VS images and IR images.

3.1 Characteristics of VS Images

The visual output from a RTO video panorama based on VS images is oriented by the visual faculties of the human eye. Color vision is a faculty that helps humans to distinguish objects from each other as it increases the contrast between them if their colors differ. The objects of interest can therefore be detected, recognized and identified easier [14]. Humans perceive colors because surrounding objects reflect electromagnetic waves that are captured by the dedicated photoreceptors on the retina (cones) if their wavelength is within the spectrum from 380 nm to 780 nm [15]. However, color vision works only at daytime or under artificial light and best under good visible conditions since cones are only activated in the presence of visible light with sufficient intensity. This also applies to visual acuity, which is another faculty of the human visual system. Thanks to visual acuity, the texture of an object of interest can be perceived in detail and therefore recognized and identified. More accurately, the perceived declining size of texture elements gives us important information about the depth of scenery [16]. If it wasn't for depth perception, ATCOs could not correctly assess speed, distance or size of an object in space. In resume, the video panorama with its transmitted VS images, furnishes ATCOs almost everything he or she would see from a conventional tower. By consequent, the visual environment is one, with almost all of its advantages, they already are used to. As stated above, the perception of information through visual images works best under good light conditions because regular VS sensors only detect reflected sun- or artificial light sources. Therefore, detection, recognition and identification processes thanks to color- and depth perception are strongly altered under bad visibility conditions and even disappear in the dark.

3.2 Characteristics of IR Images

For almost eighty years [17], military institutions have been using IR sensors to detect targets even in the dark [17, 18]. As a matter of fact, IR technologies are able to detect electromagnetic waves beyond the visible spectrum. IR wavelengths reach from 780 nm to 1 mm and are therefore not visible to the human eye. Next to thermal detectors, photon detectors are amongst the most performance IR technologies [19]. More precisely, they capture the radiation of an object of interest and by interacting with electrons on the

optical sensor; an electrical output signal is generated [19]. These signals are transformed and displayed as an IR picture which humans perceive as poorly textured, black-and-white picture. As described by Planck's law, all surfaces of objects emit electromagnetic radiation with wavelengths corresponding to their temperature. For usual surrounding temperatures, the maximum of the emitted radiation has wavelengths in the IR spectral band. In contrast to VS camera sensors, which detects light reflected by the objects, IR sensors detect this self-emitting thermal radiation of surrounding objects. Therefore, warmer surfaces (e.g. engines of an aircraft, humans or birds), usually displayed brighter, can be distinguished in high contrast from cooler ones (e.g. ground, sky). This contrast based on temperature difference compared to the color based contrast in the VS image makes detection and tracking of objects easier in the IR image. As IR imaging does not need sun- or artificial light to display objects, night vision is possible and the different wavelength improves vision under bad weather conditions (e.g. snow, fog, and rain).

3.3 Workplace to Enable Fusion of VS and IR Images

As the previous paragraphs about VS and IR images already have emphasized, there are noticeable advantages of using both optical modes. The permanent availability of VS and IR camera information could help ATCOs in specific situations. However, if the information is presented separately, it also could make them deal with higher head-down times and therefore lower situation awareness or increase workload. Therefore, the first goal of INVIDEON consists in developing a demonstrator able to display VS and IR camera images simultaneously, merged into one video panorama. As a second goal, this fusion needs to be controlled by adapted input devices that are tested with end users. In addition, the integration of the PTZ function in the merged RTO environment is to be tested, as well as its associated control modalities. This paper focuses on the aims to develop a rapid prototype of such a system and gives prospective for further research within INVIDEON.

4 Methods

A user experience focused approach was the methodological framework for three explorative workshops carried out within INVIDEON. For adequate human-machine interaction (HMI) design, rapid prototyping methods were applied with the aim to provide user-centered systems. Therefore, the user's perception of a VS/IR camera merged video panorama with adapted input devices and the PTZ-control was taken into account before, during and after the prototyping processes. In this chapter, general applied methods will be described, followed by detailed methods concerning each workshop.

4.1 Participants

A total number of seven ATCOs (all male) took part in three workshops. In the first and second workshop, four ATCOs joined per workshop; in the third workshop, three ATCOs took part. Three ATCOs were present at two workshops; one ATCO participated in all three workshops. Their professional responsibilities included runway and ground control on regional airports. They participated voluntarily and were recruited by DFS Aviation Services, a INVIDEON partner.

4.2 Material

Input Device Material. For workshop 1, three input devices to control the PTZ camera were provided: a 3D-mouse, eye-tracking glasses and a touch input device via tablet. The 3D-mouse is a device that allows ATCOs to control the PTZ camera in a tridimensional manner. More specifically, ATCOs can click on an object of interest, increase its size by a zoom function gradually or stepwise on different levels and track it manually. Thanks to eye-tracking-glasses, the PTZ-camera can be controlled by the captured eye fixations and nodding. Reflecting targets at the glasses' edges reflect infrared radiation back to captors attached to the RTO test platform. When the ATCO fixates an object of interest and nods, the requested object is magnified on a screen. By the means of a touch input device via tablet, ATCOs are able to control the PTZ via a presentation of an airport map and with the aid of a miniature panorama of the exterior view on a tablet. Some areas of interest are tagged on the map. They can be selected by tapping on the tablet; by consequent the PTZ automatically focuses on these hotspots. Furthermore, the size of objects of interest can be increased. Independent from the input device, the PTZ-video was displayed on a separate monitor and not yet included in the video panorama.

For workshop 2, a 3D-mouse to control the VS/IR fusion was provided. Thanks to this input device, ATCOs can control gradually or stepwise to which extent the video-panorama is displayed in the IR, fully merged or VS range.

Image-and Video-Material. For workshop 1, singular IR and VS video streams as well as singular IR and VS images and two merged VS/IR images (an image closer to VS and one in "pseudo-colors") were at disposition. The VS video-panorama and IR video represented scenarios from Braunschweig Wolfsburg Airport (BWE). The singular VS/IR and merged image-material was selected by project partners. Both video and image material were provided to show ATCOs the characteristics of IR and VS and to highlight their corresponding advantages. Two versions of merged VS/IR images were prepared to give an impression of how a VS/IR fusion could be displayed.

In preparation for workshop 2, several hours of traffic have been recorded simultaneously with VS and IR cameras. The videos were taken on March 7th 2017 at BWE under visual meteorological conditions by a mobile camera-carrier belonging to Rheinmetall Defence Electronics. The video material contained regular traffic (IFR & VFR) and commissioned flights (VFR) to provide a variety of elements that an ATCO normally would have to handle at a regional airport. In addition to a maximum of occurrences in a period of 20 min, other events like bird flocks for instance, were present in the selected

scenario. For workshop 2, a fully merged IR and VS panorama was provided by Fraunhofer IOSB (cf. Fig. 1.).

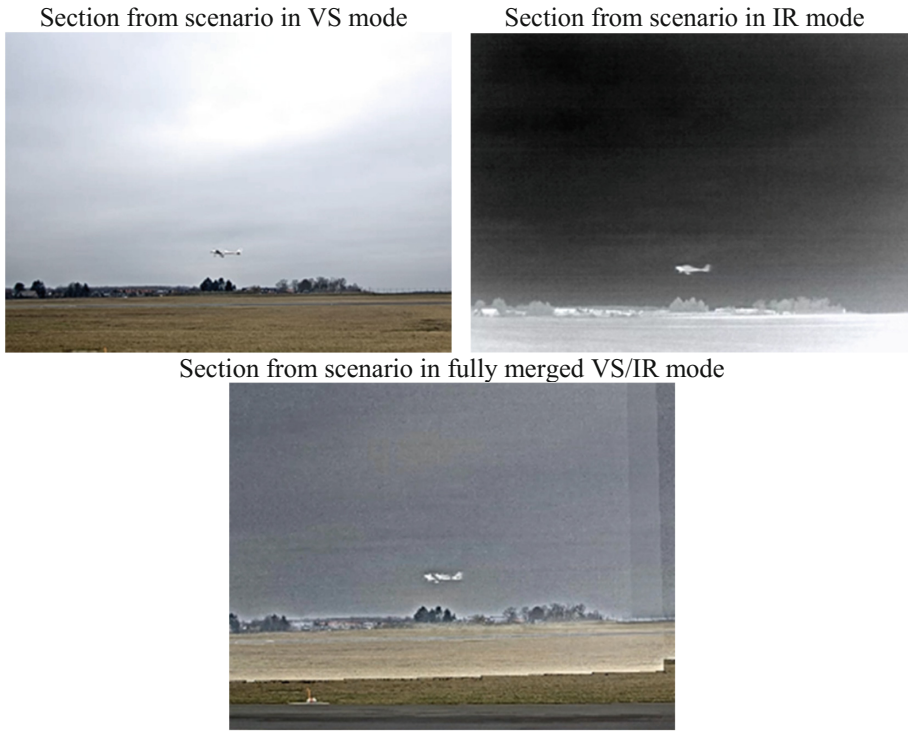


Fig. 1. VS and IR mode in a fully merged version.

Simulation Material. In preparation for workshop 1, and 3, different scenarios were created on the simulation platform at DLR. The content of each scenario was created step by step based on the project goals and the ATCOs feedback in the previous workshop.

For workshop 3, a rapid prototype (cf. Fig. 2) was created relying on the feedback and findings in workshop 1 and 2. A head-up display of the PTZ and the VS/IR merged video-panorama represent the output core prototype. The platform design relied essentially on feedback and findings from workshop 1 and 2. Therefore, a chart with integrated hotspots and a 3D-mouse inspired digital PTZ-control- and a digital slide bar to control the overlay were the basis of the ATCO's control monitor (cf. Fig. 3).



Fig. 2. Prototype of ATCO workplace in workshop 3

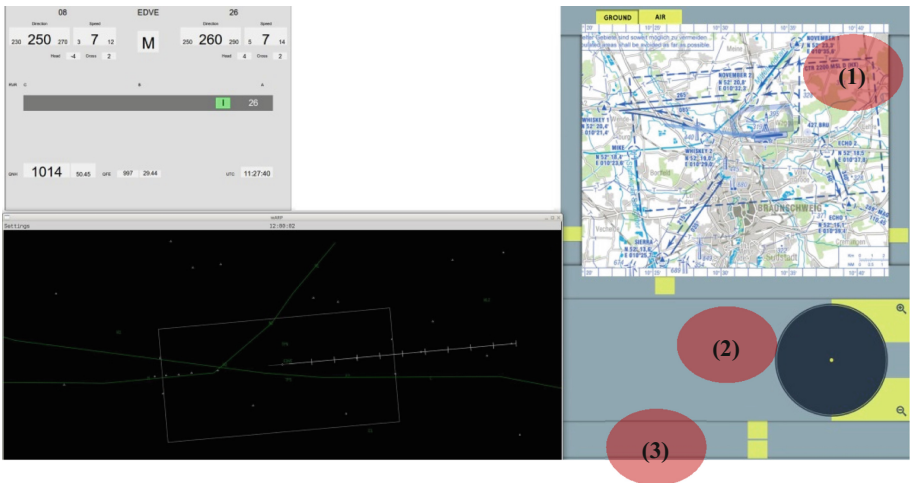


Fig. 3. Control monitor with interactive chart for PTZ function (1), digital 3D-mouse inspired PTZ-control input device (2) and slide control input device for VS/IR overlay (3)

Data Collection Material. The data collection was based on qualitative methods such as active brainstorming, open discussions and semi-directed interviews. Furthermore, data collection by quantitative methods was applied through the system usability scale [20]. A mixed approach of qualitative and quantitative methods represented the use of an adapted Cooper-Harper scale.

4.3 Workshop 1

Goals of Workshop 1. The first goal of workshop 1 consisted in presenting singular VS and IR video-streams as well as singular VS and IR images and two different merged

VS/IR images to get ATCOs' feedback about the perceived advantages and disadvantages of VS and IR modes as well as their first impression on merged VS/IR material. Secondly, workshop 1 aimed for testing three different input devices to control the PTZ function, integrated in a video panorama.

Procedure of Workshop 1. In part one, ATCOs evaluated singular VS/IR video streams as well as singular VS/IR images and two differently merged VS/IR images. In cooperation with human factors specialists, ATCOs were invited to compare both display modes and to point out advantages and disadvantages of each video mode, in relation to their daily ATC practice. Furthermore, they were asked to give feedback on the two differently merged VS/IR images.

In part two, ATCOs used the simulation platform at DLR to test three different PTZ control input devices by means of a prepared traffic scenario at visual meteorological conditions. The input devices were a 3D-mouse, eye-tracking glasses and a touch input device via tablet. Only one input device was tested per scenario. After each run, ATCOs completed a SUS-questionnaire [20] evaluating firstly the utility and usability of the tested input modality on a 5-point Likert scale (1 = totally disagree to 5 = totally agree). At the end of the questionnaire, they were asked about advantages, disadvantages, possible improvement and supplementary comments they associated with the tested input device. At the end of all three runs, ATCOs were debriefed and interviewed about their experiences with the different input devices during the experiment.

4.4 Workshop 2

Goals of Workshop 2. Workshop 2 focused at presenting a fully merged VS and IR video stream to ATCOs to receive their feedback from an operational point of view on advantages, disadvantages and possible improvement measures of the VS/IR control device.

Procedure of Workshop 2. The DLR simulation platform was used to show the fully merged video-stream from a real-time traffic scenario described in Fig. 1. ATCOs were recalled the advantages of both VS and IR modes and asked to manually control the fusion degree of VS/IR with a 3D-mouse, depending on the visual cues they would like to detect and to recognize. Thanks to the 3D-mouse, ATCOs were able to switch smoothly in gradual steps from IR to VS by turning the input device or to make bigger progressive steps by tapping on it. While the participants were watching the video and tested the VS/IR control features, the experimenter encouraged to change the display mode between VS and IR at specific events in the video (e.g. grey plane in front of grey sky). Thus, all ATCOs saw the same situation in both modes of presentation as well as in different fusion degrees. At the end of the scenario, the experimenter asked ATCOs in a semi-guided interview questions about their opinion on object detection, weather and light, input modalities and usability.

4.5 Workshop 3

Goals of Workshop 3. Workshop 3 aimed at testing the elements elaborated in the previous workshops combined in one prototype. This set-up includes a head-up PTZ camera display controlled by a 3D-mouse inspired digital input device and a slide bar for VS/IR fusion control. Feedback on the tested prototype from ATCOs should be provided to project partners so that they could, as a result, adapt it better to the operator's needs.

Procedure of Workshop 3. Two ATCOs participated in the study at the prototype test platform at the same time. One had the role to execute ATCO relevant tasks while the other one was an expert observer. Each ATCO performed both roles. The complete exercise run took two hours in total. ATCOs began by a 30 min training session which was followed by traffic scenarios under CAVOK conditions, foggy conditions and night vision; each scenario took 30 min. During the exercise run, the expert observer completed adapted Cooper-Harper scales to estimate the traffic situation management depending on visibility conditions, the use of VS/IR fusion tools and PTZ control. After each run, the active ATCO completed a SASHA [22] questionnaire where they could rate their perceived situation awareness on a 7-point Likert scale from (1 = totally disagree to 7 = totally agree) as well as on utility and usability (SUS) of the previously tested system. In a debriefing phase, ATCOs could add comments, opinions and further suggestions on the exercise and the setting.

5 Results

Due to the low number of participants, the recorded data was analyzed descriptively. In the following chapters, the results will be described separately for each workshop.

5.1 Results of Workshop 1

Feedback on Singular VS/IR Images and Merged VS/IR Images. The feedback on singular VS/IR video streams as well as on the singular VS/IR images showed that ATCOs perceived the difference of information they got from each display mode. The idea of having access to additional visual cues through IR overlay in lower visibility conditions was perceived positively. From an operational point of view, the ATCOs pointed out requirements to prevent loss of reality, false interpretations (e.g. jetwash that looks like fire in IR) and liability questions.

Concerning the two differently merged VS/IR images, they preferred the version that was closer to VS mode than the one which relied on "pseudo colors".

Test of PTZ Control Input Devices. Concerning the 3D-mouse, a total score of 61 out of maximal 100 was attained. According to Bangor et al. [21] this score indicates that the utility and usability of this device was rated as "ok". ATCOs stated that they appreciated especially the intuitive handling of the 3D-mouse but criticized the latency

in system reaction by the manual object tracking, which could result in increased head-down times.

The eye-tracking glasses achieved a total score of 51 which suggests a rather poor utility and usability performance. Even though ATCOs were very fond of the idea of not having to control the PTZ manually, disadvantages from a practical perspective emerged. Thus, ATCOs criticized that nodding was rather cumbersome and that the eye-tracking-function was not as accurate as expected. Above all, ATCOs claimed that the glasses were not comfortable to wear. Concerning improvement feedback, they endorsed the idea of an exact eye-tracking instrument without glasses that magnifies an object of interest by other means.

Concerning the touch input via tablet, a score of 43 was attained, which also stands for a poor performance in terms of perceived utility and usability. Despite the positive aspect of having a good overview on hotspots, ATCOs criticized the amount of hand movements necessary to execute the PTZ. Furthermore, the fact of not being able to swipe over the touchpad but having to tap constantly was perceived as an obstacle for active objective tracking.

In resume, the touch input via tablet scored lowest ($N = 4$; $SD = .48$) after the eye-tracking-glasses ($N = 4$; $SD = .43$). The perceived utility and usability was rated highest for the 3D-mouse ($N = 4$; $SD = .6$) (cf. Fig. 4).

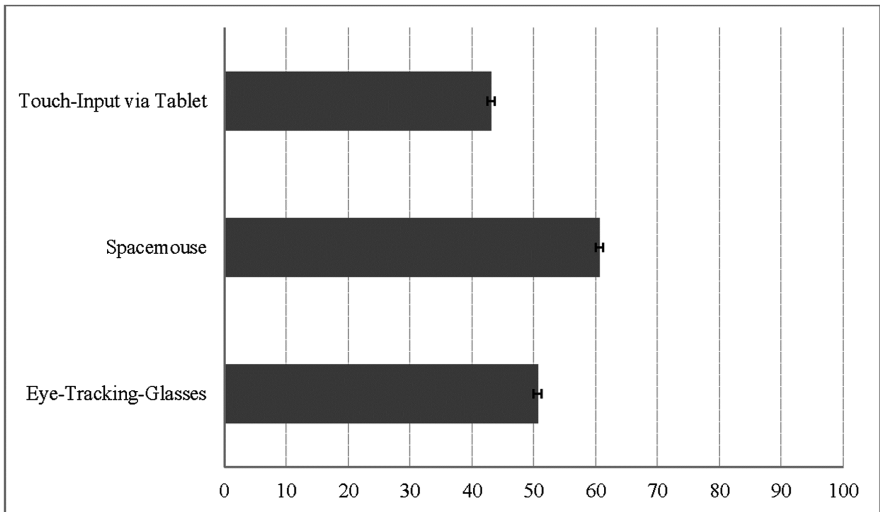


Fig. 4. SUS-Score of tested PTZ control input devices

5.2 Results of Workshop 2

The results of workshop 2 show a variety of first impressions on the fully merged video stream.

Generally, ATCOs hesitated to take clear positions on aircraft identification, correct assessment of speed, acceleration and heading of an aircraft. Moreover, they expressed that they would like to see other visibility conditions such as night- and fog scenarios.

Concerning the input VS/IR control device, one positive aspect was the possibility to “jump” from IR over fixed overlay degrees to VS. Nevertheless, others preferred the gradual movement they could apply to smoothly overlay an IR range with VS range. As an additional result, the suggestion to replace the 3D-mouse by a digital slide control device emerged.

5.3 Results of Workshop 3

The results of workshop 3 will be described in terms of situation awareness, perceived utility and usability as well as in terms of estimated traffic situation management.

Concerning the perceived situation awareness, ATCOs achieved the highest score in the foggy conditions ($N = 3$; $SD = .1$), followed by the night condition ($N = 3$; $SD = .48$) and the CAVOK condition ($N = 3$; $SD = .25$) (cf. Fig. 5). The average means of perceived situation awareness are good (CAVOK) up to very good (night and fog condition).

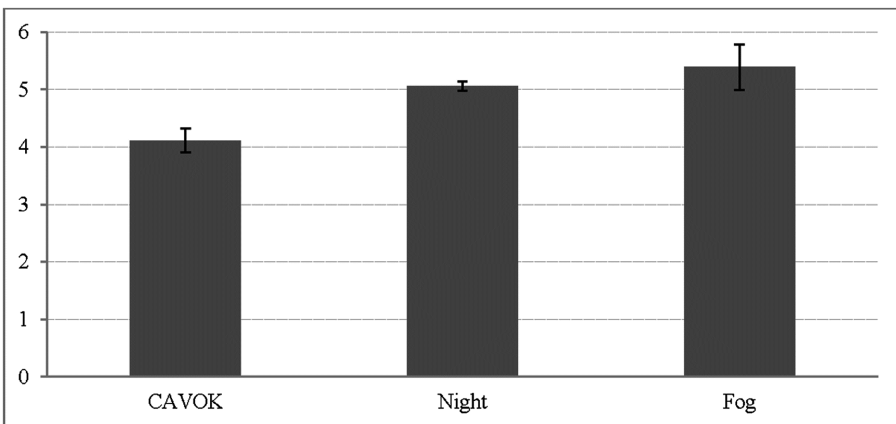


Fig. 5. Perceived situation awareness per experimental condition

The perceived utility and usability attained a mean score of 85. According to Bangor et al. [21], this score indicates that the utility and usability of the system was rated as “excellent”.

Results show that no major impairment was perceived by the observer while the active ATCO performed ATC relevant tasks and operated the VS/IR fusion and PTZ control function.

6 Discussion

Throughout the three workshops, a RTO prototype equipped with augmented reality features such as VS/IR fusion and head-up PTZ display with adapted input devices was developed with a user-centered approach. From a general and abstract concept, human factors specialists, project partners and end users worked out a concept that gradually improved.

Beginning by discovering the advantages of VS and IR modes in ATC, concrete ideas were developed in the first two workshops to redefine requirements. In fact, presenting more information to the ATCOs than they perceive currently under restricted visibility conditions would influence their work methods. Certainly, they could work on a rather constant workload if air traffic does not decrease due to bad weather conditions. Nevertheless, it has to be clarified what happens in terms of communication and liability when ATCOs see more than pilots. Concerning the VS/IR fusion modalities, ATCOs had a clear preference for a merged image that is closer to what they see in VS. However, preferring a more “realistic image” is not surprising considering the ATCOs work methods.

Regarding the perceived situation awareness in workshop 3, ATCOs rated their perceived situation awareness highest in the foggy condition ($N = 3$; $SD = .1$), followed by the night condition ($N = 3$; $SD = .48$). The CAVOK condition ($N = 3$; $SD = .25$) scored lowest but still as “good”. These results can be explained in two different ways. It is therefore possible that ATCOs detected objects of interest better due to the predominant use of IR which results in higher contrast perception due to sharp-edged contours. Another explanation is a training effect. Thus, ATCOs were already better trained in the fog and night condition compared to the CAVOK condition which was the first condition after training.

Compared to the utility and usability perception of the tested PTZ control input devices in workshop 1, the perceived utility and usability of the final prototype increased to “excellent”. The relatively high score can be explained by the results and the progress throughout the three workshops, but it has also to be considered that the user-centered approach might have an impact on the results. Integrating the final users’ suggestions into the construction cycle and adapting the object in creation to their specific needs is fundamental for successful HMI-design. Such approach is predictive for higher user acceptance and satisfaction. As past research suggests, letting final users participate in change processes reduces their resistance to change [23]. Therefore, it should be continued to include ATCOs into future studies and workshops. In this case, the created prototype could inspire ATCOs for further implementation strategies. In a next step of INVIDEON, the concept will be tested by means of live video material. Especially when integrating new features into a RTO environment that imply VS/IR video fusion, it is necessary to test them with real videos rather than in a simulation environment only. Another planned activity is to develop automatic IR-tracking as an ATCO assistance extension to PTZ object following.

References

1. Fürstenau, N.: Introduction and overview. In: Fürstenau, N. (ed.) *Virtual and Remote Control Tower*, pp. 5–12. Springer International Publishing Switzerland, Cham (ZG) (2016). https://doi.org/10.1007/978-3-319-28719-5_1
2. SAAB: Remote tower revolutionises air traffic management. SAAB (2017). <http://saabgroup.com/Media/stories/stories-listing/2017-02/remote-tower-revolutionises-air-traffic-management/>
3. Manske, P.G., Schier, S.L.: Visual scanning in an air traffic control tower—a simulation study. *Procedia Manuf.* **3**, 3274–3279 (2015)
4. Ellis, S.R., Adelstein, B.D., Reisman, R.J., Schmidt-Ott, J.R., Gips, J., Krozel, J., Cohen, M.: Augmented reality in a simulated tower environment: effect of field of view on aircraft detection (2002)
5. Ellis, S.R.: Towards determination of visual requirements for augmented reality displays and virtual environments for the airport tower. NATO R&T Organization (2006)
6. Pinska, E.: An investigation of the head-up time at tower and ground control positions. In: *Proceedings of 5th Eurocontrol Innovative Research Workshop*, pp. 81–86 (2006)
7. Schmidt, M., Rudolph, M., Papenfuss, A., Friedrich, M., Möhlenbrink, C., Kaltenhäuser, S., Fürstenau, N.: Remote airport traffic control center with augmented vision video panorama. In: *IEEE/AIAA 28th Digital Avionics Systems Conference, DASC 2009*, p. 4-E. IEEE, October 2009
8. Reisman, R., Brown, D.: Design of augmented reality tools for air traffic control towers. In: *6th AIAA Aviation Technology, Integration and Operations Conference (ATIO)*, p. 7713, September 2006
9. Roberts, D., Menozzi, A., Cook, J., Sherrill, T., Snarski, S., Russler, P., Clipp, B., Karl, R., Wenger, E., Bennett, M., Mauger, J.: Testing and evaluation of a wearable augmented reality system for natural outdoor environments. In: *Head-and Helmet-Mounted Displays XVIII: Design and Applications*, vol. 8735, p. 87350A. International Society for Optics and Photonics, May 2013
10. Hofmann, T., König, C., Bruder, R., Bergner, J.: How to reduce workload—augmented reality to ease the work of air traffic controllers. *Work* **41**(Suppl. 1), 1168–1173 (2012)
11. Fürstenau, N., Schmidt, M., Rudolph, M., Möhlenbrink, C., Halle, W.: Augmented vision videopanorama system for remote airport tower operation. In: Grant, I. (ed.) *Proceedings of 26 International Congress of the Aeronautical Sciences (ICAS)*. Optimage Ltd., Edinburgh (2008)
12. Fürstenau, N., Rudolph, M., Schmidt, M., Lorenz, B., Albrecht, T.: On the use of transparent rear projection screens to reduce head-down time in the air-traffic control tower. In: Vincenzi, D.A., Mouloua, M., Hancock, P.A. (eds.) *Human Performance, Situation Awareness and Automation Technology: Current Research and Trends*, pp. 195–200. Lawrence Erlbaum, Mahwa (2004)
13. Papenfuss, A., Friedrich, M.: Head up only—a design concept to enable multiple remote tower operations. In: *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pp. 1–10. IEEE, September 2016
14. Tanaka, J.W., Presnell, L.M.: Color diagnosticity in object recognition. *Percept. Psychophys.* **61**(6), 1140–1153 (1999)
15. Alleysson, D.: *Le traitement du signal chromatique dans la rétine: Un modèle de base pour la perception humaine des couleurs*. Doctoral dissertation, Université Joseph-Fourier-Grenoble I (1999)

16. Bajcsy, R., Lieberman, L.: Texture gradient as a depth cue. *Comput. Graph. Image Process.* **5**(1), 52–67 (1976)
17. Hudson, R.D., Hudson, J.W.: The military applications of remote sensing by infrared. *Proc. IEEE* **63**(1), 104–128 (1975)
18. Bernard, E., Rivière, N., Renaudat, M., Péalat, M., Zenou, E.: Active and thermal imaging performance under bad weather conditions (2014)
19. Rogalski, A.: Infrared detectors: an overview. *Infrared Phys. Technol.* **43**(3–5), 187–210 (2002)
20. Brooke, J.: SUS-a quick and dirty usability scale. *Usabil. Eval. Ind.* **189**(194), 4–7 (1996)
21. Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: adding an adjective rating scale. *J. Usabil. Stud.* **4**(3), 114–123 (2009)
22. Dehn, D.M.: Assessing the impact of automation on the air traffic controller: the shape questionnaires. *Air Traffic Control Q.* **16**(2), 127–146 (2008)
23. Coch, L., French Jr., J.R.: Overcoming resistance to change. *Hum. Relat.* **1**(4), 512–532 (1948)



Network Re-analysis of Boeing 737 Accident at Kegworth Using Different Potential Crewing Configurations for a Single Pilot Commercial Aircraft

Don Harris^(✉)

Faculty of Engineering and Computing, Coventry University, Coventry, CV1 5FB, UK
don.harris@coventry.ac.uk

Abstract. Most aircraft manufacturers and avionics systems suppliers are developing technology for airliners that will be flown by just a single pilot. Several different configurations for such an aircraft have been proposed but most rely to some extent on ground-based support. This paper assesses various configurations using a scenario based upon the Boeing 737 accident at Kegworth in 1989. A modified AcciMap approach supplemented by further analysis using propositional networks was utilized. From such an analysis it can be seen that a single pilot can rapidly become overloaded if the information/data exchange is not mediated by ground-based assistance. However, some configurations using ground-based support to a single pilot may also offer the opportunity to reduce communication error.

Keywords: Single pilot operation · AcciMap methodology
Propositional networks

1 Introduction

All major airlines now control operations a network control center. These go under a number of names (Operations Control Center; Flight Operations Center; Airline Operations Center) but all perform essentially the same functions. The International Air Transport Association [1] suggests that Operations Control has three basic components. These all have various sub-functions within them:

- Operations Control
 - Operations Management (including airline systems, fleet)
 - Dispatch Management (including aircraft routing/re-routing; flight and load planning; fuel planning; flight following and meteorology)
 - Maintenance Management (including maintenance control; technical specialists; aircraft on the ground; in-flight technical issues)
 - Crew Management (including crew scheduling and tracking at main operating base and downline and airport operations)

- Service Components
 - Customer (including passenger service and reservations)
 - Financial control
 - Other (including cargo handling and Hotel reservations and transportation)
- Support Components
 - Operational Coordination (including with airport management; air traffic control/management; news watch for geo-political instability, significant environmental events, etc.)
 - Operational Liaison (including Chief Pilot and base representatives)
 - Operational Support (security and safety)
 - Data Management and Analysis (delays, costs etc.).

Such centers operate 24 h/day and may employ just a few people undertaking all these functions, to several hundred personnel all performing one dedicated function. This depends on the size of the airline and the complexity of its operation. The major carriers will often have engineers from the aircraft and/or engine manufacturers embedded in these operating rooms to provide specialist technical support. Rolls-Royce, the engine manufacturer has recently opened its own dedicated engine services Airline Aircraft Availability Centre. From this facility it can monitor remotely aircraft using the latest generation of engines providing real-time support to pilots (if needed) and coordinating maintenance and repair world-wide. Indeed, this center has access to more information concerning the health and performance of the engines than do the pilots.

The objective of providing ground support from network control centers is to provide a fully integrated, multi-disciplinary support team to the pilots alleviating them of the mundane flight planning paperwork and providing them support during high-workload, non-normal and emergency operations. Providing a range of dedicated expertise should enable better decisions and minimize delays. Furthermore, ground-based monitoring should help to anticipate and pro-actively manage the impact of unplanned events.

Most aircraft manufacturers and avionics systems suppliers are developing technology for airliners that will be flown by just a single pilot. Embraer announced that it was hoping to provide single-pilot capabilities by 2020. Paul Eremenko, Chief Technology Officer has also openly stated that Airbus is developing technologies that will allow a single pilot to operate a commercial airliner. In support of the same objective, Boeing is planning to undertake initial experimental flights in 2018 where autonomous systems will take over some of the pilot's decisions. In the UK work is being undertaken as part of the Open Flight Deck program to determine the technology requirements and optimal crewing strategies for a single crew airliner.

Several different high-level configurations for a single crew aircraft have been proposed [2–4] but all solutions rely, to a greater or lesser extent on ground-based support depending upon how much onboard automation (or autonomy) is proposed. Any ground-based support may or may not be provided in real-time during flight. Comerford *et al.* [4] outlined five basic, high-level configurations for a single pilot aircraft:

1. One pilot on board, who inherits the duties of the second pilot.
2. One pilot on board, with automation replacing the second pilot.
3. One pilot on board, with a ground-based team member replacing the second pilot.

4. One pilot on board, with onboard personnel as back-ups.
5. One pilot on board, with support of a distributed team.

A future single pilot aircraft is just one part of a wider operating system with several discrete components and functions within it, such as:

- The aircraft itself, including:
 - the pilot
 - onboard automation/autonomous systems
- Ground-based component including (but not limited to):
 - ‘Second pilot’ support station/office (or ‘super-dispatcher’ – see Bilimoria *et al.*, [5])
 - Real-time engineering function
 - Navigation/flight planning function (including meteorology).

It can be seen that most of the functions and information required to support single pilot operations are already available in airline network control centers. The issue becomes one of how this support can be made available to the flight deck in a timely, optimal manner?

Irrespective of how the ground-based component is arranged, all configurations are essentially a problem in distributed cognition, and more specifically, Distributed Situation Awareness – DSA (see Stanton *et al.* [6, 7]. Different parts (human or machine actors) in the wider system hold different components of information and represent different views on the system depending upon their goals (which should be compatible but not necessarily the same). For DSA to occur there must be communication between the agents in the system (which may take many forms). Finally, one component in the system (human or machine) can compensate for degradation in Situational Awareness in another agent.

The concept of DSA operates at a system level, not at the individual level. It implies different but compatible, requirements and purposes and the appropriate information/knowledge relating to the task and the environment changes as the situation develops [8]. DSA can be represented by propositional networks [7]. Propositional networks, comprise ‘subject’ (noun), ‘relationship’ (verb) and ‘object’ (noun) network structures of the knowledge required to describe a situation. For electronic/computerized systems, these may be constructed from system logic diagrams; for the representation of human knowledge objects these are constructed from cognitive interviews.

For the design of the air and ground-support components in a distributed system such as the one proposed to support a single pilot aircraft, the question becomes how should this information be distributed and represented to support DSA? What can be learned from previous accidents about how not to do it?

The AcciMap approach was developed as an analysis methodology to identify the causal factors involved in an accident or incident within a sociotechnical context. The technique represents graphically the causal factors and maps multiple contributing factors across different levels of the sociotechnical system [9, 10]. AcciMap frames the possible causal influences underpinning a sequence of events into various organizational levels. AcciMap charts depict key input and output conditions of system components and their relationships. They are not restricted to analysis within a single organizational or functional entity, which makes this approach particularly applicable for the network

of functions underpinning the operation of an aircraft. The AcciMap methodology has been adapted by authors for various applications [11].

Combining propositional networks with the concepts within the AcciMap accident analysis methodology may provide an understanding of failures of DSA across ground and air components. It is proposed that the elements within an AcciMap Analysis may be further decomposed in a semi-hierarchical manner using propositional networks. Furthermore, the approach can also be used in a pro-active manner to describe the potential inter-relationships between the elements in various high-level configurations in a single pilot air/ground system.

This paper re-analyses the Boeing 737 (G-OBME) accident at Kegworth in 1989 [12] using a modified AcciMap approach based around the standardized AcciMap methodology described by Branford et al. [11] and supplemented by further analysis using propositional networks. The accident scenario is then used as the basis for analysis for how the problem would be tackled using various configurations for the operation of a single crew aircraft.

2 Kegworth Accident, 1989

Much has previously been written about the accident at Kegworth in 1989 [12]. It is probably one of the most analyzed accidents in aviation history. However, the richness of the set of events leading up to the crash means that it bears analysis from many perspectives.

To summarize, just after leaving London Heathrow and approximately 13 min into the flight to Belfast, the pilots noticed a severe vibration as the aircraft was climbing through 28,000 feet just 20 nm to the South-East of East Midlands Airport (near Derby). This was subsequently found to be the result of a small portion of fan blade in the left-hand engine breaking off which resulted in heavy vibration, shuddering and compressor stalling which ceased after about 20 s. This was accompanied by some smoke on the flight deck. The Commander took control of the aircraft and disconnected the autopilot. As a result of the First Officer misidentifying the malfunctioning engine from misreading the engine vibration gauges (an error that was compounded by the Commander's incorrect mental model of the air-conditioning system – he believed that all the flight deck air came from the first compressor stages on the right-hand engine, hence the smoke) the right-hand engine was throttled back and subsequently shut down. At this point the engine vibrations on the left-hand (damaged) engine reduced and the smoke also began to dissipate, suggesting that the decision to shut down the right-hand engine was indeed the correct one. However, reports from passengers and cabin crew who could see evidence of fire directly in the left-hand engine and which were transmitted to the flight deck were dismissed by the Commander.

At this point the airline asked the crew to divert into East Midlands Airport which coincidentally, was also British Midland's main operating base. This involved a right-hand turn and descent to flight level (FL) 100. The Commander elected to fly the aircraft manually. During this time the First Officer was engaged with various radio calls to both the airline's main operating base and ATC. Simultaneously he was also attempting to

re-program the Flight Management Computer – FMC (unsuccessfully) for the approach into East Midland’s airport. Having failed to do this he commenced the single engine approach checklist but was interrupted on several occasions by various radio calls from both ATC and British Midland’s maintenance facility. There was some attempt to review the situation but his was compromised by the high workload on the flight deck and the various interruptions. As a result of the reasonably tight turn and a high workload descent to land at the nearby airport, the Captain was required to extend his flight path further to the South of East Midland’s airport to increase the distance to the threshold.

The initial part of the descent and approach was normal (in the circumstances) and it was not until the landing gear was deployed and the power on the damaged engine was increased to compensate for the drag from the flaps and gear, that problems began to occur. About 2.4 nautical miles from touchdown there was an abrupt decrease in power as the left-hand engine failed completely. This was accompanied by fire warnings. The First Officer attempted to re-light the (undamaged) right-hand engine but there was not enough time and no procedure available. The aircraft crashed on an embankment of the M1 motorway near the village of Kegworth just 900 m short of the runway threshold. Forty-seven passengers were killed and 74 seriously injured.

3 Modified AcciMap Analysis of Events

A modified version of the standardized AcciMap methodology described by Branford et al. [11] was utilized for the initial analysis of events in the Kegworth accident. The standardized AcciMap model considers events at three levels prior to the final outcome:

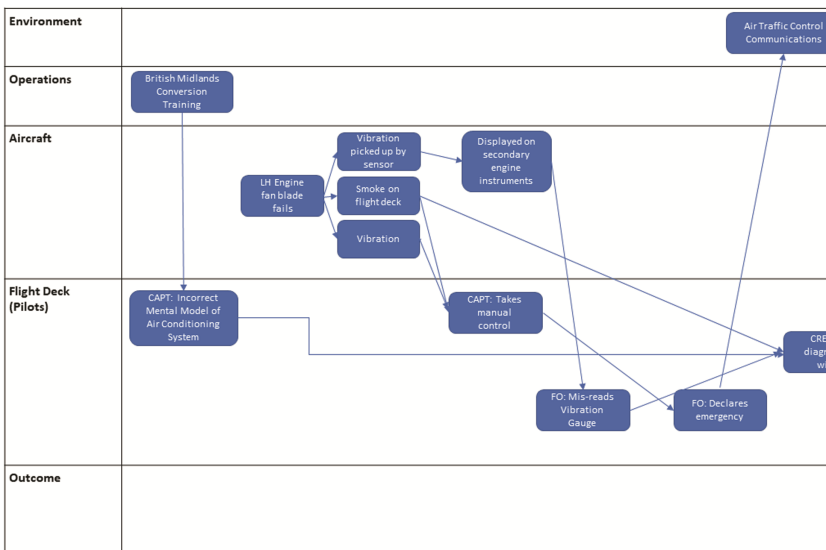


Fig. 1. Initial section of the AcciMap analysis (adopting the modified method based upon Branford et al. [11]) describing the sequence of events leading up the Boeing 737-400 accident (G-OBME) at Kegworth, 1989 [12].

External, Organizational and Physical/Actor Events, Processes & Conditions. In the further modification used in this analysis the latter level is broken down into two further sub-levels: Aircraft and Aircrew, representing the avionics and pilots, respectively. The complete AcciMap analysis is rather large. A section of the analysis of the sequence of events leading up to the accident is presented in Fig. 1. The arrows depict causal (or contributory) relationships between factors, hence an arrow from one factor to another indicates that the former was necessary for the latter to occur [11].

Described another way, the crew were faced with two basic problems:

- What is wrong with the aeroplane (and by implication, what does this mean for the management of the flight) and
- Where are we and where are we going have (and by implication how do we get there)?

Different parts of the overall system (both on aircraft and off-aircraft) held different pieces of information and represented them from different perspectives. A high-level view of the navigation problem in the Kegworth accident is described this manner in Fig. 2. This also begins to demonstrate the importance of communication between system elements for the development of DSA. One of the problems with the use of propositional networks when used to describe DSA is that there is an implicit assumption that transfer of data/information (communication) between actors is complete and accurate. However, any transmission of data/information between interfaces and users and/or person-to-person may not be perfect, especially if the interface is poor or the transmitter or receiver is under pressure. As a result, the links between agents in Fig. 2 have been adapted to include a representation of the quality of data exchange.

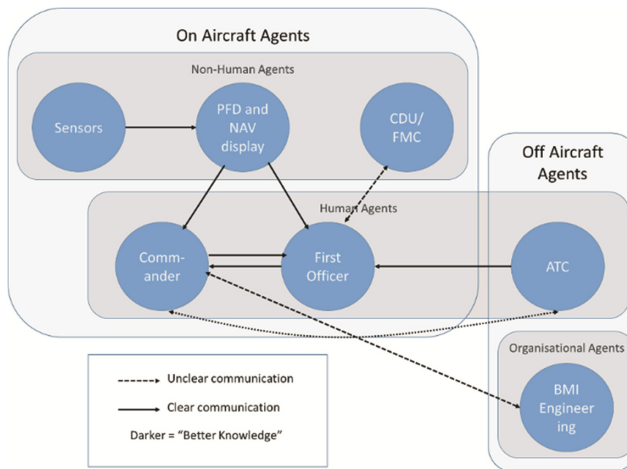


Fig. 2. A high-level view of the navigation problem in the Kegworth accident describing the communication linkages.

Some of the elements within the position problem can be further described in more detail as a propositional network (see Fig. 3) which in this case has been delineated to make it clear which physical parts of the system contained which properties.

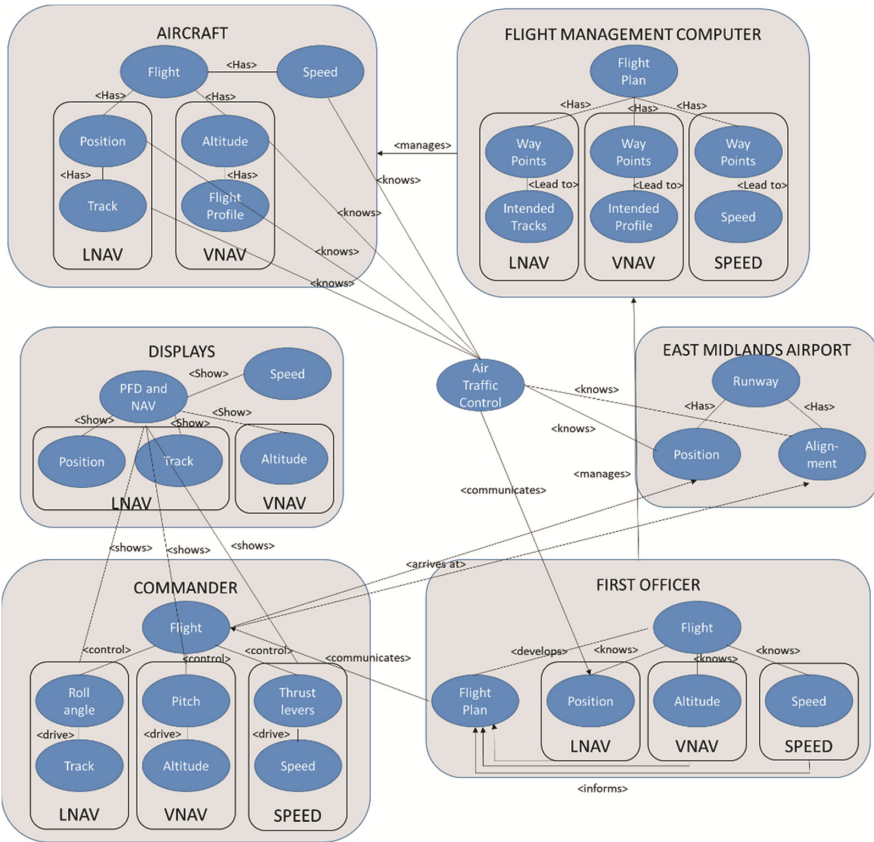


Fig. 3. Propositional network describing the navigation problem distributed across the human and non-human actors.

From a consideration of the material contained in Figs. 1, 2 and 3 it can be seen that various pieces of data/information about the problem (the engine malfunction) and the solution (managing the engine problem and navigating safely to East Midlands airport) were held in a number of on- and off-aircraft locations and by both human and non-human elements. Briefly, the engine sensors and ‘knew’ which engine was malfunctioning (but not why – they contained good data but little information) and the symptoms were displayed on the secondary engine instruments (however these were not communicated adequately as a result of the poor interface – again more emphasis on data rather than information) hence the poor awareness of the crew. Air Traffic Control had a strategic view of the position and track of the aircraft relative to East Midlands airport (and other conflicting traffic) – good information – and an idea of the crews’ intentions. The aircraft’s FMS ‘knew’ the position and orientation of the aircraft relative to the airport (data) but could not communicate it nor enact it. The navigation intent was formulated by British Midland maintenance (land at East Midlands airport) and was shared by ATC and both pilots (but not the FMS). The intent was enacted cooperatively by the First

Officer receiving vectors from ATC (tactical data) which were communicated to the Captain who was flying the aircraft manually. The First Officer had a limited tactical awareness of the navigation solution (disparate pieces of data); the Captain’s navigational awareness was even more limited, essentially restricted to the immediate altitude, course and speed communicated by the First Officer.

As the events preceding the accident progressed various knowledge objects were activated or de-activated (e.g. see Stewart et al. [8]) but when human agents were involved (either as transmitter or receiver of data/information) the quality of the data/information passed may not have been perfect. It can be seen that by representing the various actors at play in the Kegworth accident at the various levels in the AcciMap hierarchy, and by including the lines of data/information transmission, it becomes apparent that no one entity had a complete view of the situation. Situation awareness was not only distributed, it was inefficient and incomplete. Furthermore, the nexus of the communication activity (the First Officer) became overloaded, especially when dealing with the navigation problem when diverting to East Midlands airport (see Fig. 2).

It can be seen from the adapted AcciMaps in Figs. 4, 5 and 6 that when considering the navigation problem derived from the Kegworth accident scenario, the single pilot can rapidly become overloaded if the information/data exchange is not mediated by ground-based assistance (c.f. the role of CAPCOM – capsule communications – in NASA mission control). This is particularly the case when the inputs from a wider distributed team are considered (configuration 5 – Fig. 6). It can also be seen from the communication networks described in Fig. 4 that some potential modes of miscommunication (and hence error) are actually reduced. For example, single crew aircraft configuration 1 where there is just a single pilot on board. Having a ground-based pilot serves to alleviate some of the workload experienced by the pilot but only if they can

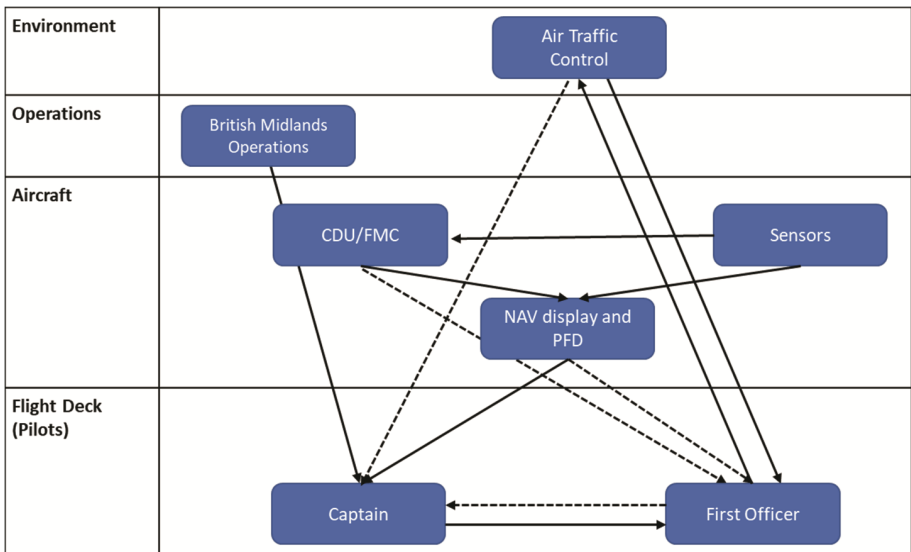
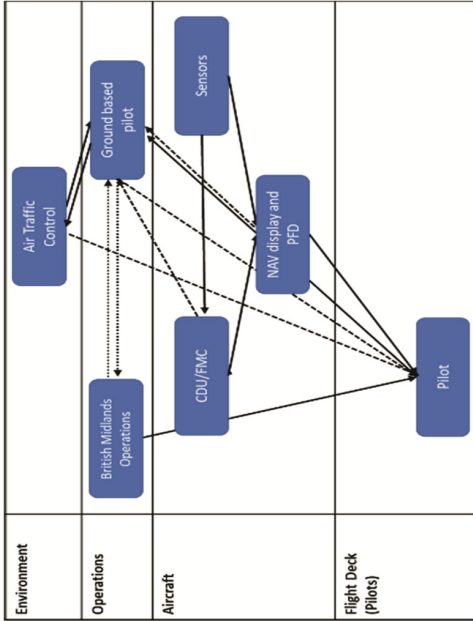
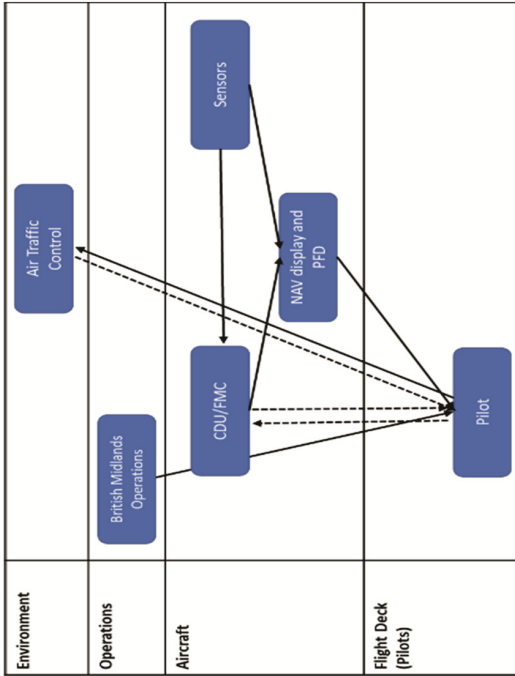


Fig. 4. Baseline navigation problem faced by the crew at Kegworth

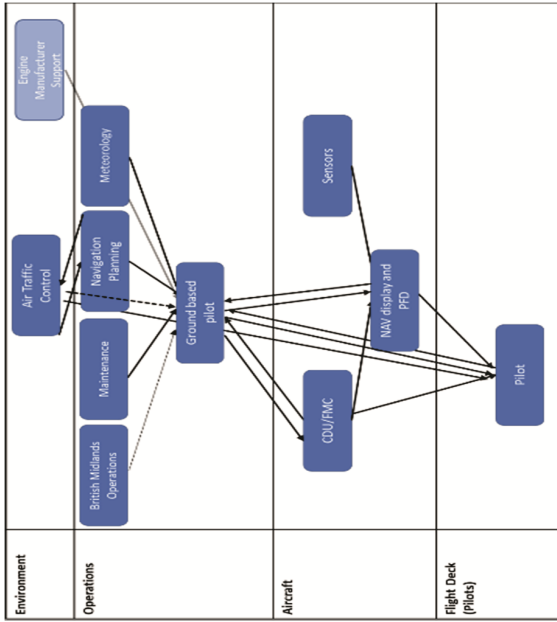


Navigation problem from the Kegworth scenario with a single pilot on board but with ground-based team member replacing the second pilot (single crew aircraft configuration 3)

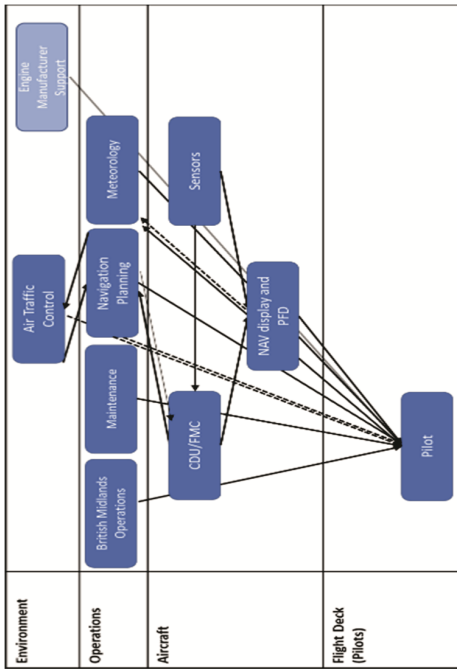


Navigation problem from the Kegworth scenario with just a single pilot on board (single crew aircraft configuration 1)

Fig. 5. Kegworth accident scenario re-described using single crew configurations 1 and 3.



Navigation problem from the Kegworth scenario with a single pilot on board but with assistance from a distributed team and a ground-based pilot (adapted version of single crew aircraft configuration 5)



Navigation problem from the Kegworth scenario with a single pilot on board but with assistance from a distributed team (single crew aircraft configuration 5)

Fig. 6. Kegworth accident scenario re-described using single crew configuration 5 and a modified version of configuration 5.

be coordinated effectively (configuration 3). The greater the number of entities in the distributed system, the more critical the role of this function becomes.

References

1. International Air Transport Association: Operations Control Centre (2014). <http://www.iata.org/whatwedo/workgroups/documents/acc-2014-gva/occ-3-occ.pdf>. Accessed 23 Feb 2018
2. Harris, D.: A human-centred design agenda for the development of a single crew operated commercial aircraft. *Aircr. Eng. Aerosp. Technol.* **79**(5), 518–526 (2007)
3. Comerford, D., Brandt, S.L., Lachter, J., Wu, S.-C., Mogford, R., Battiste, V., Johnson, W.W.: NASA's Single-Pilot Operations Technical Interchange Meeting: Proceedings and Findings (NASA/CP—2013–216513). National Aeronautics and Space Administration, Ames Research Center Moffett Field, CA (2013)
4. Stanton, N.A., Harris, D., Starr, A.: Modelling and analysis of single pilot operations in commercial aviation. In: Proceedings of HCI Aero 2014, 30 July – 1 August, 2014, Santa Clara, CA, USA (2014)
5. Bilimoria, K.D., Johnson, W.W., Schutte, P.C.: Conceptual framework for single pilot operations. In: Proceedings of HCI Aero 2014, 30 July – 1 August, 2014, Santa Clara, CA, USA (2014)
6. Stanton, N.A., Baber, C., Walker, G.H., Salmon, P., Green, D.: Toward a theory of agent-based systemic situational awareness. In: Vincenzi, D.A., Mouloua, M., Hancock, P.A. (eds.) Proceedings of 2nd Human Performance, Situation Awareness and Automation Conference (HPSAAII), Daytona Beach, FL, 22–25 March 2004
7. Stanton, N.A., Stewart, R.J., Baber, C., Harris, D., Houghton, R.J., McMaster, R., Salmon, P., Hoyle, G., Walker, G., Young, M.S., Linsell, M., Dymott, R.: Distributed situational awareness in dynamic systems: theoretical development and application of an ergonomics methodology. *Ergonomics* **49**, 1288–1311 (2006)
8. Stewart, R.J., Stanton, N.A., Harris, D., Baber, C., Salmon, P., Mock, M., Tatlock, K., Wells, L., Kay, A.: Distributed situational awareness in an airborne warning and control aircraft: application of a novel ergonomics methodology. *Cogn. Technol. Work* **10**, 221–229 (2007)
9. Rasmussen, J.: Risk management in a dynamic society: a modelling problem. *Saf. Sci.* **27**, 183–213 (1997)
10. Svedung, I., Rasmussen, J.: Graphic representation of accident scenarios: mapping system structure and the causation of accidents. *Saf. Sci.* **40**, 397–417 (2002)
11. Branford, K., Naikar N., Hopkins, A.: Guidelines for AcciMap analysis. In: Hopkins, A. (ed.) Learning from High Reliability Organisations, pp. 193–212. CCH, Sydney (2009)
12. Air Accidents Investigation Branch of the Department of Transport: Report No: 4/1990. Report on the Accident to Boeing 737–400, G-OBME, near Kegworth, Leicestershire on 8 January 1989. HMSO, London (1990)



Human Performance Assessment of Multiple Remote Tower Operations Simultaneous Take-Off and Landing at Two Airports

Peter Kearney¹, Wen-Chin Li²(✉), and Graham Braithwaite²

¹ ATM Operations and Strategy, Irish Aviation Authority, Dublin 2, Ireland

² Safety and Accident Investigation Centre, Cranfield University, Bedford, UK
wenchin.li@cranfield.ac.uk

Abstract. Remote Airport Traffic Control Centre describes the goal of providing aerodrome control service at more than one airport from a geographically separated remote control centre. The technology is applicable to all airports, regardless of size or movement rate, in some cases potentially as a primary tower and in others as a fully functioning contingency or backup system. The innovative concept of multiple remote tower operations (MRTO) can maximize cost savings by using advanced technologies at the remote tower working positions, which permits less controllers to accomplish the same quantity and quality of air traffic management tasks at an airport. The aim of this research is to assess human performance on multiple remote tower operations by using the Human Error Template (HET). The results of this research demonstrate that advanced technology based on human-centered design has improved ATCO's performance in monitoring and controlling more aircraft from two different airports. OTW design permits the adjustment of the size of view of the selected airports (100%, 75%, 50% or 25%) based on ATCO's preference, but they are also able to zoom-in by PTZ to enhance visual searching. Furthermore, OTW allows different colors to distinguish different airports, green for Cork and red for Shannon, further increasing ATCO's situation awareness to which airport he/she is engaging. The EFS system integrates aircraft strip information with the map of runway and taxiway, providing the ATCO a clear picture of the locations of the moving targets. This is a very effective design to prevent runway incursions. The information presented by the RDP can facilitate ATCO predicting the flow of traffic and landing time at each airport, thereby facilitating enhanced decision making in respect of simultaneous movements at both airports. These new technology enhancements significantly increase ATCO task performance and in conjunction with a good human centered design the ATCO's decision making capability can also be enhanced.

Keywords: Air traffic control · Human-computer interaction
Human Error Template · Multiple remote tower operations
Situation awareness

1 Introduction

The development of advanced technology in the aviation industry has significantly changed the traditional air traffic management (ATM) system and air traffic controller (ATCO) task performance. The innovative concept of multiple remote tower operations (MRTO) can improve operational safety and maximize cost savings by using video panorama-based remote tower systems which permit less controllers to accomplish the same quantity and quality of air traffic management tasks at an airport. In Europe, the Single European Sky initiative has been set up 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]. A novel solution to fulfill these objectives is for a single air traffic controller to deliver control services to multiple airports from a remote location. The application of advanced technology suggests that air traffic controllers can visually supervise aircraft and airports from remote locations by video-link from a remote tower center (RTC). It is clear that visual features of detection, recognition, and identification by MRTO can fit the requirements by regulators and air navigation service providers (ANSP's) [2]. As the Remote Tower Concept was being researched for over 20 years, it became clear that it would differ fundamentally from traditional modes of local tower operation. Cameras and sensors could be placed anywhere on the field, and ATCO's would be presented a virtual picture of reality, enhanced by a number of advanced technical devices such as panoramic digital reconstruction with high resolution pan-tilt zoom (PTZ), and electronic flight strip (EFS). The advanced design of MRTO created some human-computer interaction (HCI) safety concerns, as this system expected one ATCO to perform four ATCOs' tasks with the assistance of new technology [4].

Air Traffic growth in recent years has highlighted deficiencies in infrastructure and airspace capacities resulting in increasing delays to aircraft and passengers. In order to address these concerns, the Single European Sky initiative has been set up with the following aims - improve safety, reduce airspace user costs and minimize environmental impact whilst at the same time increasing efficiency and capacity in order to meet the requirements of growing air traffic numbers [3]. ATCOs' visual search whether in a radar centre or an aerodrome control tower is critical for maintaining SA, but can be heavily influenced by the surrounding environment and equipment interface design. In order to gain a comprehensive understanding of the effects of different HCI design on cognitive function, it is necessary to apply a holistic approach which provides a comprehensive assessment of the impact on performance [5]. The HCI design of the Controller Working Position (CWP) including Electric Flight Strips (EFS), Radar Data Processing (RDP) and Out of Window screens (OTW) impacts an ATCO's cognitive processes in terms of attention distribution, situation awareness (SA) and decision-making. There can be three ATCOs in a traditional tower including the approach controller, air movement controller and surface movements controller. The multiple remote tower operation offers the goal of providing aerodrome control services for two or more airports from a RTC without direct presence at the airports under control. The aim of multiple remote tower operation is to deliver benefits in line with SESAR's high-level objectives, to enhance ATCO's situation awareness, to improve productivity, and to enhance system contingency and reduce

workload [1]. It is based on the assumption that new technology will better facilitate ATCO's situation awareness and quality of decision-making, therefore, one controller would be able to perform all the tasks of monitoring, supervising, and communications involved in controlling aircraft and vehicles at two different airports which in a traditional system would be performed by four ATCOs. Air traffic controllers must make a rapid judgment of the situation that is being presented by their respective ATM system, and then take appropriate decisions to ensure aviation safety and maximize airspace and runway efficiency. Managing complicated ATM systems to maintain safe separation of aircraft is not only an issue of technical skill performance but also of real-time decision-making involving situation awareness and risk management within a time-limited environment [6].

The Human Error Template (HET) has been developed specifically for the aerospace industry in response to Certification Specification (CS) 25.1302. In particular, it is intended as an aid for the early identification of design induced errors, and as a formal method to demonstrate the human factors issues in the design and certification process in aviation [8]. The method consists of a checklist approach and comes in the form of an error template. HET works as a simple checklist and is applied to each bottom level task step in a hierarchical task analysis (HTA). The technique works by indicating which of the HET error modes are credible for each task step, based upon the judgement of the participants. The participant simply applies each of the HET error modes to the task step in question and determines whether any of the modes produce any credible errors [9]. The HET error taxonomy consists of twelve error modes shown below:

1. Fail to execute
2. Task execution incomplete
3. Task executed in the wrong direction
4. Wrong task executed
5. Task repeated
6. Task executed on the wrong interface element
7. Task executed too early
8. Task executed too late
9. Task executed too much
10. Task executed too little
11. Misread Information
12. Other

For each credible error the participant provides a description of the form that the error would take, such as, 'Scanning Shannon airport thinking it is Cork airport'. Next, the participant has to determine the consequence associated with the error e.g. incomplete scan of the Runway at Cork, instantaneous to scan runway at Shannon. Finally, the participant then has to determine the likelihood of the error (low, medium or high) and the criticality of the error (low, medium or high). If the error is given a high rating for both likelihood and criticality, the aspect of the interface involved in the task step is then rated as a 'concern' requiring intervention. 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 very quick to use. The HET method is also easily auditable as it

comes in the form of an error proforma. The only real disadvantage associated with HET is that for large tasks, it may become laborious to perform [10].

2 Method

2.1 Participants

Five subject-matter experts participated in six focus group sessions. The subject matter experts ages ranged between 41 and 53 year old ($M = 47.2$, $SD = 4.5$); the working experience as qualified ATCO is between 13 and 25 years ($M = 17$, $SD = 5.9$).

2.2 Apparatus

The Remote Tower Centre is equipped with 2 Remote Tower Modules (Fig. 1) comprising of 15 screens in each (14 active & 1 spare). Each of the modules is equipped with the SAAB Electronic Flight Strip (EFS) and Radar Data Processing (RDP) display which is used only as a distance to touch down indication and not to provide a Radar Service. Each of the modules accommodates 2 controller positions, Surface Movements Control (SMC) and Air Movements Control (AMC). The SAAB Remote Tower camera system was installed at Shannon and Cork Remote Tower Sites. The Cameras are located at suitable positions to provide the exact same viewing aspect as the current physical Tower at each location. The out the window (OTW) visualization is made up of 15 full HD displays in a 220° configuration. 14 displays are normally used to present the images from the 14 cameras, while the last display is a stand-by unit in the event of equipment failure. The displays match the camera resolution of 1920×1080 pixels, and have a refresh rate of 60 Hz.

2.3 Scenario

ATCO controls a Boeing-737 landing at Shannon airport whilst simultaneously controlling another Boeing-737 departing from Cork airport from the RTC situated 120 miles away in Dublin airport.

2.4 Research Design

All participants were supplied with a training package for the HTA and HET methodology which consisted of a description of the method, a copy of the methods associated error taxonomy, a flowchart showing how to conduct an analysis using the method, an example of an analysis carried out using the method, and an example output of the method. Participants were also given a HTA describing the action stages involved when remotely controlling a B737 Aircraft landing at Shannon airport. The participants were also provided with access to the MRTO module located at Dublin airport to remotely control Shannon and Cork airports. Five subject-matter experts familiar with multiple remote tower operations and human performance participated in this research. Participants had also participated in 50 trials of MRTO to gain practical experience of using the system. The Hierarchical Task Analysis (HTA) method is used



Fig. 1. The Module of multiple remote tower control system comprised by Electronic Flight Strip (EFS), Out of the Window (OTW), Radar Data Processing (RDP), Information Data Processing (IDP) and Voice Communication System

to break down activities, scenarios, and tasks into single separate operations. This methodology enables a comprehensive step-by-step description of the task activities associated with the scenario described above [11]. The step by step breakdown of multiple remote tower operations included ATCO's operational behavior and their interaction with the various pieces of equipment in the MRTO such as EFS, OTW, RDP, and IDP during which time their task performance was noted. The operational action related to HCI on multiple remote tower operations included time to complete tasks and sub-tasks which were then analyzed using the twelve error modes of HET.

Finally, participants had to determine the likelihood of the error (low, medium or high) and the criticality of the error (low, medium or high). If the error is given a high rating for both likelihood and criticality, the aspect of the HCI involved in the task step is then rated as a 'concern', meaning that it requires attention in order to assure and improve safety. The errors associated with a specific task were classed as 'Pass' or 'Concern' [8]. The definition of Pass was assigned to errors whose effects would not endanger the safety of MRTO operations (scores between 1 and 4). Conversely, the Concern rating applied to errors where there was a high probability of occurrence and their safety criticality was also high (scores between 6 and 9). 'Concern' highlighted HCI design issues which could lead to critical human factors accidents/incidents; which should prompt the designer/regulator to consider changes to, or redesign of, interfaces, procedures, and/or ATCO's training, in order to avoid these errors presenting in multiple remote tower operations (Fig. 2).

| | | | | |
|------------|-------------|----------|-------------|-----------|
| | | Low 1 | Medium 2 | High 3 |
| Likelihood | Low 1 | 1 | 2 | 3 |
| | Medium 2 | 2 | 4 | 6 |
| | High 3 | 3 | 6 | 9 |
| | Criticality | | | |

Fig. 2. The HET likelihood and criticality matrix with the Pass and Concern respectively highlighted in green and red (Color figure online)

3 Results and Discussions

The application of HET was based on the HTA to analyze step-by-step of multiple remote tower operations (Cork and Shannon airports). This permitted an accurate assessment of the actions and the cost of the effort and time required to complete the operational steps, such as checking the RDP to estimate the distance and timing of arriving flights, monitoring moving aircraft/vehicles on the aerodrome by OTW, or inputting information into EFS. The objective was to understand the limitations of human performance and human-computer interactions on multiple remote tower operations. Once the overall task goal of performing multiple remote tower operations has been specified, the next step is to break the overall goal down into meaningful sub-goals [11]. In the task, “simultaneous aircraft Landing at Shannon airport and Departing aircraft plus a Circuit at Cork airport”, the overall goal was broken down into sub-goals. All of these operational actions have to be assessed based on twelve error modes of HET to identify the design induced error related to HCI on MRTO by all participants. The example of HET evaluation form shown as Table 1.

The results of this research demonstrate that advanced technology based on human-centered design has improved ATCO’s performance in monitoring and controlling more aircraft from two different airports [12]. OTW design permits the adjustment of the size of the percentage of the selected airports (100%, 75%, 50% or 25%) based on ATCO’s preference, but they are also to zoom-in by PTZ to enhance visual searching. Furthermore, OTW allows different colors to distinguish different airports, green for Cork and red for Shannon, further increasing ATCO’s situation awareness to which airport he/she is engaging. The EFS system integrates aircraft strip information with the map of runway and taxiway, providing the ATCO a clear picture of the locations of the moving targets. If an ATCO has permitted one aircraft to move toward the runway, he/she will not be able to permit another aircraft moving to the same runway with this EFS. This is a very effective design to prevent runway incursions. The information presented by the RDP can facilitate ATCO predicting the flow of traffic and landing time at each airport, thereby facilitating enhanced decision

Table 1. Example of HET output for predicting HCI design induced error on multiple remote tower operation

| Scenario: simultaneously landing on EINN and departing on EICK | | | Task step: 1.2.4 scan of EINN OTW + HDP (5 s) | | | | | | | | |
|--|-------|---|--|------------|---|---|-------------|---|---|------|---------|
| Error mode | TIC K | Description | Outcome | Likelihood | | | Criticality | | | PASS | CONCERN |
| | | | | H | M | L | H | M | L | | |
| Fail to execute | V | No check on EINN | Possible runway incursion | | V | | | V | | V | |
| Task execution incomplete | V | Incomplete scan of the runway | Possible runway incursion | | V | | V | | | | V |
| Task executed in wrong direction | | | | | | | | | | | |
| Wrong task executed | V | Scanning Cork thinking it is Shannon | Possible runway incursion | | V | | | V | | V | |
| Task repeated | V | Repeated scan of EINN | Time consuming | | | V | | | V | V | |
| Task executed on wrong interface element | V | Scanning Cork thinking it is Shannon | Possible runway incursion | | V | | | V | | V | |
| Task executed too early | V | Scanning of Shannon is done at an early stage | Increased workload as subsequent scans will be carried out | | | V | | | V | V | |
| Task executed too late | V | Scanning of Shannon is done at a later stage | Delayed situational awareness | | | V | | | V | V | |
| Task executed too much | V | Repeated scan of EINN | Time consuming | | | V | | | V | V | |
| Task executed too little | V | Incomplete scan of the runway | Possible runway incursion | | V | | | V | | V | |
| Misread information | V | Scanning without paying attention | Possible runway incursion | | V | | V | | | | V |
| Other (extra unexpected calls) | | ...if increasing workload, the likelihood of certain error modes may increase as well | Depending on the type, feed in turn into the criticality of the error... | | V | | | V | | | |

making in respect of simultaneous movements at both airports. These new technology enhancements significantly increase ATCO task performance and in conjunction with a good human centered design the ATCO's decision making capability can also be enhanced. Therefore, it is possible for one ATCO to perform the tasks originally designed for four ATCOs' to complete.

There are also some potential HCI risks to be aware of [6]. The aim of multiple remote tower operation is to deliver benefits in line with SESAR's high-level objectives. It is based on the assumption that new technology will facilitate ATCO's situation awareness and quality of decision-making, therefore, one controller would be able to perform all the tasks of monitoring, supervising, and communicating activities involved in controlling two different airports [1]. The results of this research based on 50 field trials and scientific research framework demonstrated that MRTO is a safe approach to control both air and ground movement for two low volume airports simultaneously whilst maintaining safety. However, how much is too much when it comes to tasks for a single ATCO? MRTO is safe whilst operations are normal, the evolution of a critical event at one or other of the airports has the potential to overload the single ATCO, this requires additional study and analysis before MRTO operations can be deployed.

Only two error modes raised safety concerns with HET for MRTO in this research, both task executed incomplete and misread of information on the operational step of 1.2.4 Scan of EINN OTW and RDP in five seconds. Though the majority of operational steps are marked as PASS with medium likelihood and low criticality (Table 2), such as task repeat on scan EINN runway was time consuming, task executed too late leading to lack of situation awareness, these do increase ATCO workload as the steps are required to be repeated to assure safety. Furthermore, the 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 unusual pilot request there is potential for workload to increase and time pressure to become more acute. 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 a very quick method to use. The error taxonomy used is also comprehensive; it was based on existing error taxonomies from a large number of human error identification (HEI) methods [9].

Table 2. Summary of HET on simultaneously controlling aircraft landing and departing based on multiple remote tower operations

| Error Modes (below numbers shown as %) | | Fail to execute | Task execution incomplete | Task executed in wrong direction | Wrong task executed | Task repeated | Task executed on wrong interface element | Task executed too early | Task executed too late | Task executed too much | Task executed too little | Misread information | Other |
|---|---|-----------------|---------------------------|----------------------------------|---------------------|---------------|--|-------------------------|------------------------|------------------------|--------------------------|---------------------|-------|
| | | | | | | | | | | | | | |
| 1 | Answer phone to coordination call from EINN APP (10-15 seconds) | 60 | 100 | 40 | 40 | 60 | 0 | 20 | 20 | 0 | 0 | 60 | 0 |
| 2 | Insert strip into ARR sequence on EFS (3 seconds) | 60 | 80 | 20 | 60 | 40 | 0 | 40 | 0 | 20 | 0 | 80 | 0 |
| 3 | Check apron on OTW for push back approval on EICK(5 seconds) | 60 | 100 | 40 | 0 | 40 | 20 | 0 | 40 | 20 | 0 | 100 | 20 |
| 4 | Scan on OTW+RDP of EINN to monitor push back (5 seconds) | 40 | 60 | 20 | 60 | 40 | 0 | 60 | 20 | 0 | 40 | 60 | 0 |
| 5 | Acknowledge call + reply from EINN arrival (8 seconds) | 60 | 80 | 60 | 20 | 80 | 40 | 0 | 40 | 0 | 20 | 80 | 0 |
| 6 | Utilize OTW picture to identify A/C on approach (3 seconds) | 80 | 100 | 0 | 40 | 20 | 0 | 60 | 40 | 20 | 20 | 80 | 0 |
| 7 | Check EFS of EICK monitoring vehicles/aircraft for Taxi instruction (3 s) | 60 | 80 | 20 | 60 | 40 | 60 | 40 | 20 | 0 | 20 | 100 | 20 |
| 8 | Cross check OTW + RDP of EINN to maintain SA (3 seconds) | 80 | 80 | 0 | 60 | 20 | 0 | 60 | 20 | 40 | 0 | 60 | 0 |
| 9 | Scan runway for obstruction for landing clearance of EDNN (15 s) | 80 | 60 | 40 | 80 | 20 | 0 | 40 | 60 | 20 | 0 | 100 | 0 |
| 10 | Record clearance to land on EFS (2 seconds) | 80 | 60 | 80 | 60 | 40 | 20 | 60 | 60 | 0 | 20 | 80 | 0 |
| 11 | Line up clearance for EICK (5 seconds) | 60 | 80 | 40 | 60 | 20 | 0 | 40 | 20 | 20 | 40 | 80 | 0 |
| 12 | Move EFS on board (1 seconds) | 40 | 100 | 60 | 20 | 0 | 40 | 20 | 60 | 0 | 20 | 80 | 0 |
| 13 | Scan anemometer issue surface wind vector on EINN (3 seconds) | 80 | 80 | 80 | 20 | 60 | 40 | 0 | 80 | 40 | 20 | 100 | 20 |
| 14 | Monitor aircraft touchdown and roll on EINN runway (10-15 seconds) | 80 | 80 | 60 | 20 | 40 | 0 | 0 | 60 | 20 | 40 | 80 | 0 |
| 15 | Scan of EICK runway for take-off instruction (5 seconds) | 40 | 100 | 40 | 40 | 60 | 20 | 60 | 60 | 40 | 20 | 80 | 0 |
| 16 | Issuance take off clearance EICK (5 seconds) | 80 | 60 | 40 | 20 | 60 | 0 | 40 | 60 | 20 | 40 | 60 | 0 |
| 17 | Issue runway exit and taxi route for EDNN (8 seconds) | 80 | 80 | 60 | 80 | 0 | 0 | 20 | 40 | 20 | 60 | 80 | 0 |
| 18 | Cross check OTW + RDP on EICK for maintaining SA (2 seconds) | 60 | 100 | 20 | 60 | 60 | 20 | 40 | 60 | 20 | 20 | 100 | 0 |

The patterns of fixations on the indicators or the areas of interest (AOIs) can reveal an operator's visual trajectory of attention on the processing tasks. Eye movement patterns shown that OTW is the most important source of information to ATCO to perform his task integrated with Pan-Tilt-Zoom (PTZ). Moreover, the percentage of fixations on the relevant AOIs is deemed as the predictor of the overall SA performance. Again, the OTW is the highest percentage of fixation (76.67%). In addition, the fixation duration is the average time fixating on an AOI, which can reflect the level of importance or difficulty in extracting information. Fixation duration might reveal how long ATCOs sustain attention whilst scanning the information in order to completing the mission. Furthermore, EFS has the highest average fixation duration display. It reveals that EFS is either be the most important or the most difficult tool for managing the tasks associated with safely completing multiple remote tower operations (Fig. 3).



Fig. 3. ATCO's cognitive processing to plan the ground movement on EFS for multiple remote tower operation (EFS purple colour Shannon vs green colour Cork) (Color figure online)

4 Conclusion

Designing and managing human-computer interactions requires an understanding of the principles of cognitive systems and the allocation of functions between human operators and computer support systems. Human-centred design of multiple remote tower operations must be based on a strategic, collaborative and automated concept of operations to increase both airspace efficiency and safety [7]. The HET method is applied to determine whether or not the interface design under analysis is appropriate. The analysis assigned Pass or Concern was based on the associated error probability and criticality. The focus is on the human performance associated with the new technology in the MRTO and ensuring that the interfaces and support tools are used safely and efficiently to control aircraft both remotely and for multiple airports. The results demonstrate that advanced technology can provide sufficient technical support

to one ATCO performing a task originally designed to be performed by four ATCOs, however, the application of this new technology also induced huge workload for one ATCO. It must be stated that this research is based on normal operations and does not consider the impact of an unusual situation or critical event during the operation. Should an unexpected event occur it is likely that work will snowball thus having a negative impact on ATCO's performance through increased time pressure and increased workload. This creates a need for further research on how to manage HCI issues to increase the safety margin within MRTTO operations and provide more resilience for ATCO's cognitive abilities of decision-making. There is a need for further research on how to manage HCI issues for multiple remote tower operations to relieve ATCO's workload.

References

1. Eurocontrol: Eurocontrol Seven-Year IFR Flight Movements and Service Units Forecast: 2014–2020 (Reference No. 14/02/24-43), Brussels, Belgium (2014)
2. Fürstenau, N., Mittendorf, M., Friedrich, M.: Model-based analysis of two-alternative decision errors in a videopanorama-based remote tower work position. In: Harris, D. (ed.) EPCE 2014. LNCS (LNAI), vol. 8532, pp. 143–154. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07515-0_15
3. Eurocontrol: ATM Cost-Effectiveness (ACE) Benchmarking Report with 2014–2018 Outlook, Brussels, Belgium (2015)
4. Hollan, J., Hutchins, E., Kirsh, D.: Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.* **7**(2), 174–196 (2010)
5. Lafond, D., Champagne, J., Hervet, G., Rousseau, R.: Decision analysis using policy capturing and process tracing techniques in a simulated naval air-defence task. *Hum. Factors Ergon. Soc. Annu. Meet.* **53**(18), 1220–1224 (2009)
6. Ltifi, H., Kolski, C., Ayed, M.B.: Combination of cognitive and HCI modeling for the design of KDD-based DSS used in dynamic situations. *Decis. Support Syst.* **78**, 51–64 (2015)
7. Schuster, W., Ochieng, W.: Performance requirements of future trajectory prediction and conflict detection and resolution tools within SESAR and NextGen: framework for the derivation and discussion. *J. Air Transp. Manag.* **35**, 92–101 (2014)
8. Stanton, N.A., Salmon, P., Harris, D., Marshall, A., Demagalski, J., Young, M.S., Waldmann, T., Dekker, S.: Predicting pilot error on the flight deck: validation of a new methodology and a multiple methods and analysts approach to enhancing error prediction sensitivity. Elsevier (2008)
9. Stanton, N.A., Harris, D., Salmon, P.M., Demagalski, J., Marshall, A., Waldmann, T., Dekker, S., Young, M.S.: Predicting design-induced error in the cockpit. *J. Aeronaut. Astronaut. Aviat.* **42**(1), 1–10 (2010)
10. 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)
11. 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)
12. Kearney, P., Li, W.-C., Lin, J.: The impact of alerting design on air traffic controllers' response to conflict detection and resolution. *Int. J. Ind. Ergon.* **56**, 51–58 (2016)



CONTACT: A Human Centered Approach of Multimodal Flight Deck Design and Evaluation

Anne-Claire Large, Cedric Bach^(✉), and Guillaume Calvet

Human Factors and Ergonomics Department,
Bertin Technologies, Toulouse, France
{anne-claire.large, cedric.bach,
guillaume.calvet}@bertin.fr

Abstract. This paper proposes a review of the latest theories in cognitive sciences about the multimodal nature of cognition. Based on this state of the art the paper introduces a dedicated user centered design and evaluation process for multimodal flight deck. This process is called CONTACT meaning *COckpit NaTural interACTion*. The entire process is described step by step from the analysis of *Needs* for a new multimodal interaction project till the identification of the *multimodal solutions*, passing by the description of pilot's *intentions*, the *perceptivo-motor experience* and *capabilities* and the *modalities selection*. This process is illustrated by some examples and lessons learned. A set of improvements of the method is also provided, in order to mature this new approach of multimodal interaction design for flight deck, which has been applied on projects for aeronautical industry.

Keywords: Human Factors · Methods · Multimodal interaction
User centered design · Flight deck

1 Introduction

The past decades introduced a new law in Human Computer Interaction sciences, beyond the Moore's law, we are talking about the Buxton's *Law of Promised Functionality* [5]. The Buxton's law illustrates the exponential growth of interaction means including multimodal interaction. Therefore raises a new complexity for the user centered designer looking for a tighter coupling between multimodal interaction possibilities and Human skills and previous experiences. Actually, the HCI popular *Human-computer metaphor* shows some limitations when the designer expects multimodal interactions compatible with Human skills; also called "Natural interactions". Therefore this paper expects to present a summary of the latest multimodal theories in cognitive sciences and an approach of multimodal interactions in line with these theories. In this way, this paper proposes a user centered design process, focusing on the design of the future flight deck, dedicated to the identification of the most compatible multimodal interactions according to the piloting tasks. So the paper starts with a state of the art of the latest multimodal theories in cognitive science. Then the paper

describes a design and evaluation process for multimodal interaction, called CONTACT (for COckpit NATural interACTION), in line with these theories. Finally the paper concludes on lessons learned and way forwards.

2 Theoretical Background

Since the 1950s Cognitive Sciences have been dominated by the cognitivist (or computo-symbolic) approach and the Human-computer metaphor. According to this way of thinking, the mental activity is based on computations made on symbols, in step-by-step processing. Representations from modal systems are transduced into amodal symbols that represent knowledge. Moreover, the perception (considered as an input data) and the motoric response (considered as an output) are functionally dissociated. Consequently, the cognitivist approach is particularly focused on computo-symbolic processing which are supposed to link those two phenomenon [11]. This dualist vision dissociates the abstract cognition on one hand and the body, the physical and the social environment on another hand. Thus, it assumes an independence of cognitive activities towards the body and the environment [2, 3, 6, 10]. Applied to ergonomics, those theories particularly focused on the cognitive processing arising between information presentation and the motoric response despite perceptive and motoric activities themselves [7].

Since a couple of decades, some new approaches defend the idea that cognition is not an amodal system but a multimodal system and that every human activity, however “abstract”, relies directly on perception and action. Those approaches, generally called embodied and situated, are more adaptive and functionalist. They consider that the sole function of cognition is the action in order to adapt to the world. The action is no longer considered as an output of the system, but as a central and essential component of cognition. Consequently, the purpose of Psychology is no longer centered on mental representations but on the dynamic body-cognition-environment interactions. Those interactions are guaranteed by the perception and the action which are the interfaces between Humans and the world to adapt to [8, 9]. In a Human Factors perspective, we consider that embodied and situated approaches should be fostered to design Human-machine interaction. This choice aims to fit with the more and more multimodal technologic conjuncture [14].

2.1 Multimodal Approaches in Cognitive Sciences: Embodied and Situated Theories of Cognition

Both terms *embodied* and *situated* are fundamentally linked but they nuance the emphasis on bodily and emotional factors on one hand [8] and on environmental, cultural and social factors on another hand [12] for studying cognition. In cognitive ergonomics, this interesting distinction allows working on different scales, from perceptive-motor details of interaction to the whole context in which it arises [4, 7].

Among recent theories of embodied and situated cognition, Barsalou’s simulationist theory [2, 3] supports that the human brain continuously simulates the possible interactions with its environment, at a perceptive, motoric and interoceptive (related to

internal states such as emotions) level. Those simulations are based on neural activation patterns previously activated during interactions with similar environments. More specifically, every interaction with the world generates a distributed neural activation on the different perceptive, motoric and interoceptive systems. Those activations are captured in associative areas in the form of a multimodal state, in a bottom-up process. The simulation process is top-down: the confrontation with an event (perceptive, motoric or interoceptive) previously experimented reactivates a multimodal state and consequently a pattern similar to a previous distributed (multimodal) activation. During each experience in the environment the brain simulates the possible interaction (including situated actions) on the basis of past interactions. Thus, every situation generates a multimodal perceptivo-motor simulation.

The ideomotor theory [11] shares many common assumptions with the simulationist approach. It suggests that the actions are represented according to their perceived effects. The initiation of an action contextually adapted would be possible only via the sensorial activation of the endogenous or exogenous effects that this action will produce [18]. Hommel et al. [11] state that the ideomotor phenomenon is based on three postulates. First, perception and action are functionally linked into the same system. Second, perception and action are represented in a distributed format, that is multimodal. Third, the action control is proactive. More specifically, the distributed (or multimodal) characteristics of experienced contexts are integrated into episodic traces in the form of event files. Those files contain multimodal perceptive and motoric information. In comparison with Barsalou's simulationist approach, the reactivation of an event file is more attributed to individual intentions instead of the encounter with the environment. Thus, the early representation of action's consequences (i.e. the goal to reach), in other words individual's intention, is sufficient to reactivate multimodal events files. The expected effects prime the action which will produce those effects, thus the action does exist before its execution.

Most recently, Versace et al. [23] proposed a memory model based on a unique, distributed and multimodal system: the *Act-In* (activation-integration) model. This model considers that the knowledge is composed of sensory, motoric, emotional and motivational properties of past experiences and that the knowledge emerges from the situation. The *Act-In* model also assumes that memory is an episodic, multimodal and distributed system. The knowledge emerges from the coupling between present experience and memory traces of past experiences. The knowledge emergence is based on two mechanisms: an inter-trace activation mechanism allows activating different memory multimodal traces containing perceptive, motoric or emotional properties common with the current situation; an intra-trace activation mechanism associates the different properties to form a trace. Both mechanisms allows simultaneously the knowledge emergence, but also the creation or modification of traces in memory. Thus, the brain is considered as a categorization system which is developing through traces accumulation and the emergence of specific knowledge depends on the singularity of memory traces and current situations, in other words their distinctiveness against other traces or situations. Interestingly, *Act-In* considers that memory traces reflect all the components of past experiences including perceptive, motoric, emotional and motivational properties, which determine the actions to a large extent. Like in simulationist and ideomotor theories, *Act-In* conceives direct links between perception and action and

defends that cognition is multimodal. Also, it supports the idea of an early activation of action in cognitive activities and rejects the idea of action as an output of the system.

According to those approaches, interacting with the world is a phenomenon emerging from the Human-environment coupling. It is a contextual phenomenon spatially and temporally situated. Consequently, it is a dynamic phenomenon. This point involves that every behavior is influenced by the previous behavior(s). For instance, Smith [19] and Thelen and Smith [20] explain the dynamic functioning of cognitive activities by the Piaget's *A-not-B* error [16]. In Piaget's experiment children between 8 and 12 months are in front of two hides (*A* and *B*). When the experimenter hides a toy behind *A* (the child see him), the child catches the toy behind *A*. But when the experimenter hides this toy behind *B* (the child still see him), the child persists to search behind *A*. According to dynamic approaches, once the child catches the object, a trace of this activity becomes an input for the following trial. Thus, the action to catch *A* emerges from the combination of this trace and the stimulus (a person hiding a toy). In other words, every situation produces a specific neural trace and a reinforcement of this trace. Behaviors are influenced by traces previously activated, that is, by prior behaviors. The most common dynamic effects are the intramodal facilitation whereby a pre-activation in one modality facilitates the following processings in the same modality (or combination of modalities) whatever the nature of the processing [21, 22] and the switching cost whereby a change of modality induces a cost [15].

2.2 Towards Applications for Multimodal Interaction

The embodied and situated approaches give strong basis about the multimodal nature of human activities. In particular, this description of human behavior indicates that the completeness of human behaviors is attached to tangible perceptive and motoric experiences. For example, if a pilot has the intention of changing its altitude, a perceptivo-motor activation associated with altitude changing will automatically be simulated at neural level (e.g. the action to perform on the rotary knob, the auditory, visual and tactile feedback of the rotary knob, the proprioceptive feedback of the plane, etc.).

In terms of application for multimodal interaction design, the first issue is to identify the perceptive and motoric modalities the most strongly related to the tasks to perform. For instance, we could imagine that the concept of fuel leak has a strong relation with the smell of fuel; thus, introducing artificially a smell of fuel in the cockpit could ease the failure detection and processing. Secondly, we have seen that the cognition is a dynamic system and that every behavior is influenced by previous behaviors. This involves adjusting the modalities allocation regarding their integration into the interaction dynamics.

In short, the multimodal interaction design has to be based on (1) the modalities allocation and (2) the interaction dynamics design. In this way, we propose to ground our method on the concept of PMU (Perceptivo-Motor Unit), presented in the following section.

The cognitive processes involved in our study are particularly automatic. Thus we expect that our method may allow quick answers, low human error and workload. Moreover, an interaction based on previous knowledge will necessarily limit the

training needs. By analogy, we can say that this approach aims to solicit skills based behaviors as described by Rasmussen [17].

2.3 Perceptivo-Motor Units (PMU)

The core idea of the embodied and situated theoretical background is that the Human behavior emerges from the encounter between the individual and its environment. This emergence takes the form of a neural activation comprising perceptive and motoric properties associated with the situation. On one hand, the individual has intentions and perceptivo-motor experiences. In addition, we consider that individual has perceptivo-motor capacities varying according to internal and environmental factors. On another hand, the environment affords stimuli and action means. We call this emerging behavior a Perceptivo-Motor Unit or PMU. The PMU is defined as a division of user’s activity making arising a behavior based on a neural trace comprising perceptive and motoric properties associated with the situation (Fig. 1).

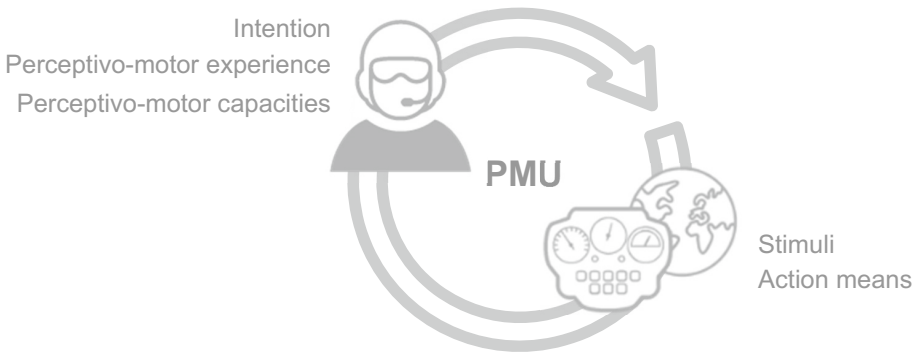


Fig. 1. Perceptivo-Motor Unit or PMU

The application of PMU concept consists in matching the elements of the individual side and the environment side in order to bring about a behavior adapted to the situation. Furthermore, the modalities allocation must pay attention to enable intramodal PMU sequences rather than switching costs (Fig. 2).

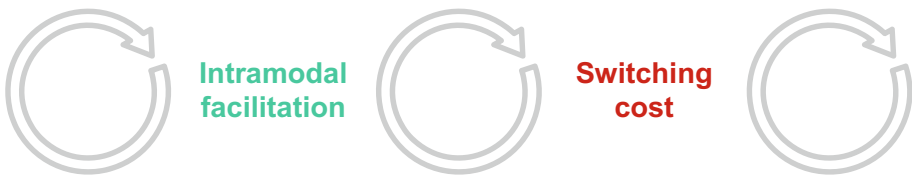


Fig. 2. Types of PMU sequences

3 Contact

3.1 Overview

CONTACT (COckpit NaTural interACTion) is a design and evaluation method for multimodal interaction projects in aeronautics environments. It is a user-centered approach based on the embodied and situated theories of cognition presented in previous sections. This method is still under development and the following section proposes a first presentation of CONTACT.

The CONTACT method has three main steps: NEEDS, CONCEPT and SOLUTION. It aims to (1) gather the needs related to the study; (2) define a multimodal interaction concept focused on perceptive and motoric aspects of interaction (Human and user centered aspects of interaction); and (3) translate the concept into technical solution (technical aspects of interaction). Those steps are built around the PMU concept derived from the literature (Fig. 3).

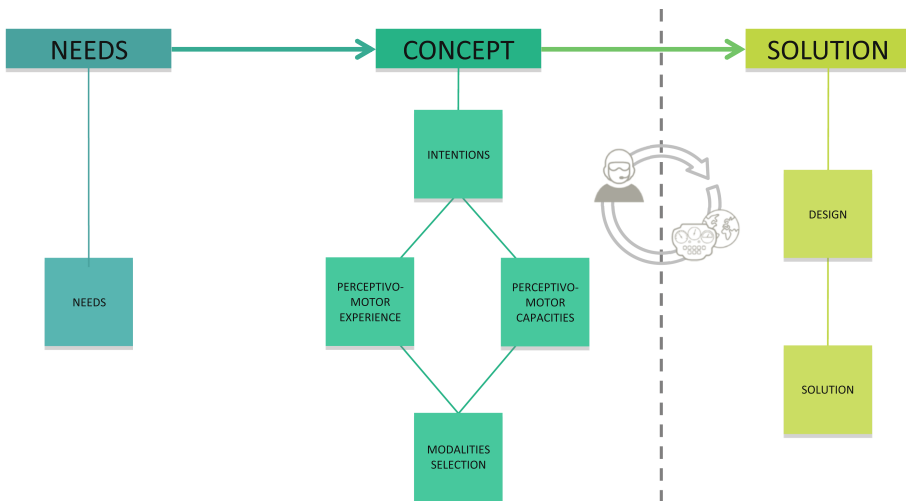


Fig. 3. CONTACT method

Each subpart of the method is presented as a document to fill. Thus, the method application is guided all along the process. Except the *Needs* document, all the documents to fill have the same template: necessary inputs, method description, example of application, expected output and evaluation to conduct. The following sections describe each step of the CONTACT method.

3.2 Needs

The *Needs document* aims to gather all the necessary inputs for the study's scope. Elements to collect concern general data (e.g. general needs, starting hypothesis,

interaction technologies assumptions, prototyping and test means), using conditions (system, users, environment and activity description) and organizational information (e.g. project stakeholders, planning). The *Needs document* offers an exhaustive view of all the inputs and documents associated to the project.

3.3 Concept

3.3.1 Intentions

Once the project needs collected, the CONTACT method proposes to determine a concept of multimodal interaction centered on the Humans aspects. As our approach is based on use cases, the first step consists in selecting relevant use cases and adapting them in the form of *intentions*. The notion of *intention* designates the expected consequences of a behavior, the goal to achieve in order to produce a perceptible effect [11]. It is the reason for which an individual performs an action. So, an *intention* includes perceptive and motoric properties associated with a goal and a situation. For example, “retract the landing gear” or “add the waypoint FJR on the flight plan” are *intentions*. This notion is close to the notion of task, but its formulation is adapted to work on perceptive and motoric properties associated with the situation.

The *Intentions document* contains criteria to select the relevant use cases and guidelines to reword tasks into *intentions*. Furthermore, the method provides examples of use cases, examples of *intentions* and associated templates to fill.

3.3.2 Perceptivo-Motor Experience

Perceptivo-motor experience refers to what future users would experience in terms of perception and action for each *intention*. This part concerns exclusively what they would experience “ideally”. It doesn’t take in account the constraints associated with the activity such as turbulence, weather events, incapacitation, abnormal situations, etc. This step aims to gather as much information as possible about future user’s knowledge (in the field of aeronautics, mainstream technologies or everyday life in the physical world) in order to transpose this existing *perceptivo-motor experience* into the concept of interaction. The purpose of this approach is to limit training needs, errors and workload and to improve human performance in the future cockpits.

To this end, the CONTACT method provides a set of the strongest perceptive and motoric modalities based on 32 helicopter piloting tasks [13]. To create this norm, 45 participants (experts and non-experts of helicopters piloting) had to assess to what extent the tasks involved 14 motoric modalities (e.g. head, left hand, left fingers, left foot, language) and 5 perceptive modalities (vision, hearing, touch, proprioception, smelling) on Likert scales (from 0 to 5). An example of results is presented in Table 1. Using this norm consists in transposing the results from 0 to 5 to related or similar *intentions*. According to the *project needs*, this set of tasks could be completed (the protocol is provided to extend the set to further tasks or different types of aircrafts such as airplanes or UAVs). More globally, the *perceptivo-motor experience* document contains all the guidelines and templates to achieve this step. To conclude, the *perceptivo-motor experience* step gives the “ideal” modalities related to each *intention*, rated from 0 to 5 (Table 1).

Table 1. Example of perceptivo-motor experience for the intention “set the radio frequency 103.40”.

| | MOTRICITY | | | | | | | | | | PERCEPTION | | | | |
|--------------------------------|-----------|------|----------|------|-----------|--------------|------------|---------------|-----------|------------|------------|---------|-------|----------------|-----------|
| | Head | Eyes | Language | Bust | Left hand | Left fingers | Right hand | Right fingers | Left foot | Right foot | Vision | Hearing | Touch | Proprioception | Olfaction |
| Set the radio frequency 103.40 | 0,93 | 2,00 | 1,87 | 0,53 | 2,13 | 2,40 | 1,00 | 1,27 | 0,00 | 0,00 | 3,87 | 2,13 | 2,40 | 1,13 | 0,00 |

3.3.3 Perceptivo-Motor Capacities

Perceptivo-motor capacities refers to the user’s capacities to act and perceive regarding the situation’s constraints for each *intention*, with a use cases approach. This part concerns what the users could experience “realistically”. To determine those capacities, the CONTACT method provides 21 criteria impacting directly *perception* and *action* in piloting environments. The criteria are grouped into 4 categories corresponding to mission, environment, Human and cockpit characteristics. This list of criteria has been defined by experts committees and optimized to both cover all the possible impacts on *perceptivo-motor capacities* and ease its use. Nonetheless, other criteria could be added according to projects’ needs (for example, this list could be reviewed for ground control stations contexts which are substantively different from cockpits).

For each *intention*, the criteria are instantiated according to objective data. As far as possible, the criteria instantiations are graduated from 0 to 5 and this scale corresponds to precise definitions and objective data. For instance, the turbulence criterion includes different levels of turbulence defined and graduated from 0 to 5 (e.g. 5 corresponds to extreme turbulence). Once the use cases criteria are instantiated, the method allows to translate the results into perceptive and motoric modalities availability (also noted from 0 to 5). Using the previous example, high turbulence will degrade touch modality and fine motor modalities such as fingers which could be noted 1 for example. The method guides the *perceptivo-motor capacities* notation, but it does not yet provide generic rules to determine it. Thus, we recommend to involve experts of the domain for this step. As for the *perceptivo-motor experience*, this part gives the “realistic” modalities related to each intention, rated from 0 to 5 (Table 2).

3.3.4 Modalities Selection

The *modalities selection* consists in confronting *perceptivo-motor experience* (“what do I prefer doing”) and *perceptivo-motor capacities* (“what I can do”). The selection of Human modalities depends on choices between preferred modalities and available modalities. The notation from 0 to 5 used for both experience and capacities ease this choice. Indeed, a strong overlapping between two notes indicates a relevant modality. Although currently the *modalities selection* is made by experts, we are looking at automatizing this step, at least for highlighting the strong overlapping (Table 3).

Once a first *modalities selection* is done, the method proposes to check the modalities sequences fluency in order to foster intramodal facilitation and to limit switching costs. To this end, the method provides guidelines for modalities choices according to Allen’s temporal intervals [1], that is the temporal relations between the tasks. Metaphorically, this step consists in “choreographing” the interaction with a broader perspective (not focused on intentions level).

Table 2. Example of perceptivo-motor capacities for the intention “set the radio frequency 103.40” under turbulence conditions.

| | MISSION | | | | | ENVIRONMENT | | | | HUMAN | |
|--------------------------------|--------------|--------------|---------------------|---------------------------|-------------------|-------------|------------|-------------------|----------------|--------|----------------|
| | FLIGHT PHASE | FLIGHT LEVEL | HEAD DOWN / HEAD UP | FLIGHT MODE (AP / MANUAL) | NORMAL / ABNORMAL | WIND | TURBULENCE | TERRAIN DUSTINESS | CLOSE OBSTACLE | STRESS | INCAPACITATION |
| Set the radio frequency 103.40 | EN-ROUTE | 5 | 3 | AP | NORMAL | 0 | 4 | N/A | 0 | 0 | 0 |

| COCKPIT | | | | | | | | | | |
|--------------------------------|---------------------|---------------------|----------------|----------------|-----------|---------------------------|-----------------------|-------------------------|--------------------------|-------------------|
| | EXTERNAL VISIBILITY | INTERNAL VISIBILITY | EXTERNAL NOISE | AIRCRAFT NOISE | HMI NOISE | COMMUNICATIONS IN COCKPIT | VIBRATIONS IN COCKPIT | NOMINAL COCKPIT MOTIONS | UNWANTED COCKPIT MOTIONS | SMELLS IN COCKPIT |
| Set the radio frequency 103.40 | 2 | 2 | 3 | 1 | 3 | 2 | 3 | 2 | 3 | 3 |

| | MOTRICITY | | | | | | | | | | PERCEPTION | | | | |
|--------------------------------|-----------|------|----------|------|-----------|--------------|------------|---------------|-----------|------------|------------|---------|-------|----------------|-----------|
| | Head | Eyes | Language | Bust | Left hand | Left fingers | Right hand | Right fingers | Left foot | Right foot | Vision | Hearing | Touch | Proprioception | Olfaction |
| Set the radio frequency 103.40 | 2 | 2 | 3 | 1 | 3 | 2 | 3 | 2 | 3 | 3 | 3 | 2 | 1 | 2 | 5 |

Table 3. Example of modalities selection for the intention “set the radio frequency 103.40”.

| Set the radio frequency 103.40 | MOTRICITY | | | | | | | | | | PERCEPTION | | | | |
|--------------------------------|-----------|-------------|-------------|------|-------------|--------------|------------|---------------|-----------|------------|-------------|-------------|-------|----------------|-----------|
| | Head | Eyes | Language | Bust | Left hand | Left fingers | Right hand | Right fingers | Left foot | Right foot | Vision | Hearing | Touch | Proprioception | Olfaction |
| EXPERIENCE | 0,93 | 2,00 | 1,87 | 0,53 | 2,13 | 2,40 | 1,00 | 1,27 | 0,00 | 0,00 | 3,87 | 2,13 | 2,40 | 1,13 | 0,00 |
| CAPACITIES | 2 | 2 | 3 | 1 | 3 | 2 | 3 | 2 | 3 | 3 | 3 | 2 | 1 | 2 | 5 |

The *modalities selection* takes the form of an Excel document to fill, which represents in parallel the *perceptivo-motor experience*, the *perceptivo-motor capacities* and the dynamic relations between the tasks. As a way forward, we plan to automatize partially this task with a dedicated tool. Finally, the *modalities selection* step allows to select one or several relevant modalities, taking into account both embodied knowledge and situation constraints, in a Human centered point of view.

3.4 Solution

3.4.1 Design

The *design* step consists in translating the *concept* (Human aspects) into design solution. This iterative step involves several stakeholders (domain experts, developers, Human Factors specialists, designers, users, etc.) in design sessions. The design sessions’ inputs are the *needs* and the *concept* (Human modalities selected). All the previous results are presented to the stakeholders who are guided to propose several designs. The workshops are made to first propose a maximum of solution and second converge towards a few designs. Several iterative sessions could be made to choose the most efficient design(s).

Those designs could be tested during design exposures, for example while confronting future users to mockups. To this end, the CONTACT method provides an evaluation guide and usability criteria to assess the designs (e.g. error rate, task

duration, number of action, workload). The results could lead to revise or to validate the designs.

3.4.2 Solution

Finally, the designs selected are converted into technical solutions. Four levels of description have been defined to specify the technical needs: device (precise definition of the device), device hardware and software settings (e.g. mouse resolution and acceleration), interaction technic (definition of the interactions effects – e.g. click on the button starts a video and highlights the icon) and fine-tuning (interaction fine-tuning definition). This step could be made in collaboration with developers. If some prototyping and/or simulation means are available in the project, the solution could be implemented and tested in simulation conditions. The CONTACT method also provides an evaluation guide and usability criteria to assess the solution(s). Again, the results could lead to revise or to validate the solution(s).

4 CONTACT Applications Lessons Learned and Way Forward

The CONTACT method first applications (on future commercial aircrafts and future helicopters cockpits) are in progress. Those applications allows us to evaluate and improve the process and the associated tools. So far, we observed that the method using required a strong intervention of Human Factors specialists. In particular, the *perceptivo-motor capacities* and the *modalities selection* steps necessitate the involvement of Human Factors specialists. As a way forward, we are considering a partial automation of the method. Firstly, we are establishing some rules to facilitate the *perceptive-motor capacities* definition according to the set of criteria (for example, if the turbulences criteria = 4, then the touch capacities = 1). Secondly, we are studying a dedicated tool to ease the *modalities selection*. Such a tool would suggest modalities to select, in this way the users of the tool will only have to check the results obtained. Such improvements will accelerate the overall process and also allow the method using by a wider range of users (e.g. not only Human Factors specialists). Despite those future enhancements, we still recommend to convene a multidisciplinary group of specialists to optimize the benefits of the method deployment (e.g. Human Factors specialist specialized in Cognitive Sciences and physiology, engineers, UX designers, final users, developers, domain experts).

Beyond the stakeholders to involve in the project, we observed that the CONTACT method required some appropriate means. Those resources are documentation (technical documents about the targeted system, its use context, the users population, the missions to perform, etc.), but also test or simulation means to perform evaluations. According to the test and simulation means available, the CONTACT method may remain usable on subset; for example it is possible to use it without performing evaluations.

Finally, the CONTACT method can be integrated at different maturity levels of a project, upstream or during the project (e.g. from TRL 1 to TRL 6). The method is also applicable incrementally for research and development project; for example, the

CONTACT method is compatible with agile methods. The estimated duration for its application varies between a few weeks up to few month according to the project scope.

5 Conclusion

The CONTACT method interest lies in the integration of embodied and situated approaches of cognition (Human centered approach), context and interaction dynamics into an ergonomics process. In a Human Factors point of view, the expected benefits are to limit training needs, errors and workload and to improve Human performance in future cockpits. In an industrial point of view, the expected benefits are a better integration of Human Factors and safety impacts and thus an optimization of design and development cycles for complex multimodal aeronautics systems. Applying this method, also ensures a consistent modalities allocation philosophy, as they are studied at intention and also inter-intention level. To conclude, the CONTACT method is an original approach to design aeronautics multimodal workstations.

References

1. Allen, J.F.: Maintaining knowledge about temporal intervals. In: *Readings in Qualitative Reasoning About Physical Systems*, pp. 361–372 (1990)
2. Barsalou, L.W.: Perceptions of perceptual symbols. *Behav. Brain Sci.* **22**(04), 637–660 (1999)
3. Barsalou, L.W.: Grounded cognition. *Annu. Rev. Psychol.* **59**(1), 617–645 (2008)
4. Beaudouin-Lafon, M.: Designing interaction, not interfaces. In: *Proceedings of the Working Conference on Advanced Visual Interfaces*, Gallipoli, Italy, pp. 15–22. ACM (2004)
5. Buxton, W.: Less is More (More or Less). In: Denning, P. (ed.) *The Invisible Future: The seamless integration of technology in everyday life*, pp. 145–179. McGraw Hill, New York (2001)
6. Damasio, A.R.: *L'Erreur de Descartes*. Odile Jacob, Paris (1995)
7. Dourish, P.: *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press, Cambridge (2001)
8. Glenberg, A.M.: Embodiment as a unifying perspective for psychology. *Wiley Interdiscip. Rev.: Cogn. Sci.* **1**, 586–596 (2010)
9. Glenberg, A.M.: Few believe the world is flat: how embodiment is changing the scientific understanding of cognition. *Can. J. Exp. Psychol.* **69**, 165–171 (2015)
10. Glenberg, A.M., Witt, J.K., Metcalfe, J.: From the revolution to embodiment 25 years of cognitive psychology. *Perspect. Psychol. Sci.* **8**(5), 573–585 (2013)
11. Hommel, B., Müsseler, J., Aschersleben, G., Prinz, W.: The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* **24**(05), 910–926 (2001)
12. Hutchins, E.: *Cognition in the Wild*. MIT press, Cambridge (1995)
13. Lagrasta, M., Large, A.-C., Calvet, G., Brunel, L.: Interaction multimodale dans les hélicoptères du futur: Norme pour les activités motrices et perceptives associées à 32 tâches de pilotage d'hélicoptère. In: *Proceedings of the Workshop Trace*, Nanterre, France (2017)
14. Large, A.-C., Ferrari, V., Foare, H., Brouillet, D.: Donner corps à l'interaction Homme-Machine: cognition incarnée et située en ergonomie. In: *Proceedings of the Workshop Trace*, Montpellier, France (2014)

15. Pecher, D., Zeelenberg, R., Barsalou, L.W.: Verifying different-modality properties for concepts produces switching costs. *Psychol. Sci.* **14**(2), 119–124 (2003)
16. Piaget, J.: *Traité de Psychologie Expérimentale*, vol. 7. Presses universitaires, Paris (1963)
17. Rasmussen, J.: Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Trans. Syst. Man Cybern.* **13**(3), 257–266 (1983)
18. Shin, Y.K., Proctor, R.W., Capaldi, E.J.: A review of contemporary ideomotor theory. *Psychol. Bull.* **136**(6), 943–974 (2010)
19. Smith, L.B.: Cognition as a dynamic system: Principles from embodiment. *Dev. Rev.* **25**(3), 278–298 (2005)
20. Thelen, E., Smith, L.B.: *A Dynamic Systems Approach to the Development of Cognition and Action*. Bradford Book Series in Cognitive Psychology. MIT Press, Cambridge (1994)
21. Van Dantzig, S., Pecher, D., Zeelenberg, R., Barsalou, L.W.: Perceptual processing affects conceptual processing. *Cogn. Sci.* **32**, 579–590 (2008)
22. Vermeulen, N., Corneille, O., Niedenthal, P.M.: Sensory load incurs conceptual processing costs. *Cognition* **109**, 287–294 (2008)
23. Versace, R., Vallet, G.T., Riou, B., Lesourd, M., Labeye, É., Brunel, L.: Act-In: an integrated view of memory mechanisms. *J. Cogn. Psychol.* **26**(3), 280–306 (2014)



A System for Evaluating Pilot Performance Based on Flight Data

Sha Liu^{1(✉)}, Youxue Zhang², and Jintao Chen¹

¹ Flight Department, Air China, Tianjin 300300, China
13702114801@163.com

² Flight Technology College,
Civil Aviation University of China, Tianjin 300300, China

Abstract. Pilots are the most active elements in flight activities. Pilots' operation performance could affect flight safety directly. The main purpose of this study is to develop a flight operation performance evaluation system based on QAR data and a quantitative evaluation method model. In this model, one or several of the flight parameters could be selected for combination to objectively evaluate the pilot's performance of flight operations. The system was expected to be used to evaluate, analyze, pre-alarm and improve the performance of the flight operation of the pilot after one flight task or in a period of time to provide practical technical support for airlines to monitor and control flight risk. This system used a more effective method of evaluating and calculating pilot's operation performance. And the airline's performance rewards and punishments would get a more accurate and objective basis from this system.

Keywords: Flight operation · QAR data · Performance evaluation

1 Introduction

Aviation accidents have been contributed mostly by human factors. Pilots are the most active elements in flight activities, so pilots' operation performance could affect flight safety directly. Statistics by the global aviation accidents, the pilot is the key factor of the flight safety. Many results show that more than 60% of the flying accidents were caused by pilot errors, it is the main contributing factor of aviation accidents and incidents [1–3]. According to the statistics of recent years on commercial flight accidents in China, the percentage of accidents caused by the crew factors is rising. The flight crew factors contributed to 61.54% of accidents in 2001–2006 [4], and it increased to 64.58% in 2006–2015 [5]. In fact, no matter what causes the flight accident, it would eventually behave in the operation behavior [6, 7]. Therefore, it is of great practical significance to improve the pilot's operation performance and reduce the crew errors to prevent the aviation safety accident.

The flight quick access recorder (QAR) is a system that can acquire aircraft operational data easily. It includes airborne equipment for recording parameters such as speed, attitudes, altitude, control deflections, etc. A ground software station for developing software to process the QAR data and obtain the required meteorological quantities by taking into account the aerodynamic factors of the aircraft types

commonly operated by the local airlines [8]. The QAR can record all kinds of aircraft parameters, pilot operation parameters, environmental features, and alarm information during an entire flight. The practice has proved that QAR data are helpful for improving flight safety management and quality control [9]. However, the data have been rarely utilized in research.

This paper described the main features of the QAR data analysis system and illustrated its application in evaluating pilot performance. The main purpose of this study is to develop a flight operation performance studies. This evaluation system was based on flight quick access recorder data and a risk evaluation model. The system is expected to be used to store, analyze, evaluate, pre-alarm and improve the performance of the flying operation of the pilot, to provide practical technical support for airlines to monitor pilots' flight operation performance and flight risk.

2 Methodology

2.1 Quick Access Recorder Data

QAR data includes a large number of flight, operation, environmental and other types of airplane information, it mainly used in aircraft fault detection and simple operation management in current. The use of a large number of data is lack of system, and is not effective, resulting in the waste of information.

The flight QAR data, which is based on related operational rules and regulations, is used by commercial airlines to monitor and analyze the aircraft status and pilot operation performance in flight. When flight data exceeds the prescriptive normal range [8], a QAR Exceedance Event [10] or Unsafe Event is recorded by our system. The QAR Exceedance Event was divided into two levels. The first level was called Detect Limit, while the second level was called Alert Limit. Taking the Boeing 737-800 model as an example, the model had 108 types of QAR Exceedance Events, and a part of the types were shown in Fig. 1.

Exceedance Events may not lead to serious consequences. However, they could increase the probability, and could bring potential risks to aircraft and even passengers.

2.2 Evaluation Model

At present, the use of QAR data by most airlines was only limited to QAR data analysts to extract the data after flight execution for the traditional management of QAR Exceedance Events. That is, according to the level and frequency of QAR Exceedance Events standards to monitor flight and evaluate the pilots performance. Most airlines give up research on the data and ignore the value of the data. In a sense, the flight risk is ignored. A large amount of QAR data can reflect the pilot's operational characteristics at all stages of flight clearly. Taking QAR data of a flight fleet in a long period of time as the sample space, we studied the probability distribution of the entire fleet P_{fleet} according to statistical principles. With the same methodology, the probability distribution of one single pilot was calculated, and then compared with P_{fleet} , in order to evaluate and predict the pilot's risk of Exceedance Event.

| Code (i) | Description | Detect Limit | Alert Limit | Unit | Duration |
|-------------|---|-------------------|-------------------------|---------|----------|
| 100 | Excessive power during taxi in | ≥ 5 | ≥ 8 | Kts | 1sec |
| 101 | Taxi Speed: before take-off | > 30 | ≥ 40 | Kts | 3sec |
| 102 | High Taxi Speed Whilst Turning | ≥ 14 | ≥ 18 | Kts | 2sec |
| 108 | Excessive EGT (in flight) | | $\geq \text{EGT limit}$ | | 1sec |
| 109 | Rotation speed high | $\geq V_r+15$ | $\geq V_r+20$ | Kts | 1sec |
| 111 | Rotation speed low | $< V_r$ | $\leq V_r-5$ | Kts | 1sec |
| 113 | Unstick speed high | $\geq V_2+25$ | $\geq V_2+30$ | Kts | 1sec |
| 117 | Pitch attitude high during take off | ≥ 8.8 | ≥ 9.9 | Deg | |
| 119 | Pitch rate high at take off | ≥ 3.5 | ≥ 4.0 | Deg/Sec | |
| 121 | Pitch rate low at take off | ≤ 1.3 | ≤ 1.0 | Deg/Sec | |
| 123 | Climb speed high between 35 and 1000ft | $\geq V_2+30$ | $\geq V_2+35$ | Kts | 2sec |
| 127 | High roll during take off:0-35ft | ≥ 5 | ≥ 6 | Deg | 1sec |
| 131 | High roll:above 400ft/1500ft | ≥ 33 | ≥ 35 | Deg | 2sec |
| 133 | Height loss on climb below 1500ft | ≥ 30 | ≥ 100 | Ft | |
| 134 | Late landing gear retraction | ≥ 300 | ≥ 500 | Ft | |
| 141 | Flap placard speed V_{fe} | | $\geq V_{fe}$ | Kts | 2sec |
| 155 | Altitude deviation | | ≥ 250 | Ft | 2min |
| 157 | Descent rate high between 2000 and 1000ft | ≥ 1500 | ≥ 1800 | Ft/min | 3sec |
| 158 | Descent rate high between 1000 and 500ft | ≥ 1300 | ≥ 1500 | Ft/min | 3sec |
| 159 | Descent rate high below 500ft and 50ft | ≥ 1100 | ≥ 1300 | Ft/min | 2sec |
| 162 | Bank: between 500 and 200ft | ≥ 15 | ≥ 20 | Deg | 2sec |
| 163 | Bank: between 200 and 50ft | ≥ 8 | ≥ 10 | Deg | 2sec |
| 164 | Bank: below 50ft | ≥ 4 | ≥ 6 | Deg | 1sec |
| 168 | Maximum speed below 2500ft | > 230 | > 250 | Kts | 2sec |
| 169 | Approach speed high between 500 and 50ft | $\geq V_{ref}+25$ | $\geq V_{ref}+30$ | Kts | 3sec |
| 171 | Approach speed high below 50ft | $\geq V_{ref}+15$ | $\geq V_{ref}+20$ | Kts | 2sec |
| 173 | Above glideslope:1000-100FT | ≥ 1.0 | ≥ 1.3 | Dot | 2sec |
| 178 | Late land gear | ≤ 1500 | ≤ 1300 | Ft | |
| 181 | Late land flap | ≤ 1200 | ≤ 1000 | Ft | 1sec |
| 187 | Pitch attitude high at landing | ≥ 7.4 | ≥ 8.3 | Deg | 1sec |
| 189 | Pitch attitude low at landing | ≤ 1 | ≤ 0.5 | Deg | 1sec |
| 193 | Long Landing | ≥ 2500 | ≥ 3000 | Ft | |
| 195 | High normal accel (at landing) | ≥ 1.68 | ≥ 1.89 | g | 1sec |
| 197 | High normal accel (2nd bounce) | | $\geq 1.8+1.5$ | g | 1sec |

Fig. 1. QAR Exceedance Event standard sample

Quantitative evaluation method is one of the important methods for risk assessment. Generally, statistical and computational methods were used to multiply the probability of risk occurrence and the severity of its consequences to obtain the risk value. This method has less qualitative analysis, and it has higher accuracy [11, 12]. Based on the large sample statistics of QAR data, it was found that most flight performance parameters, such as touchdown distance, vertical acceleration, and pitch angle, are approximately normal distribution in large sample space ($n > 100$) [13]. Therefore, we can set a healthy fleet in a stable environment, each kind of flight

parameter distribution will be approximately a normal distribution in a long period. Then the occurrence probability of the various parameters of the aircraft fleet will also tend to be relatively stable.

The risk value is obtained by multiplying the probability of the occurrence of the risk with the severity of the consequences. As a result, the severity of each pilot QAR Exceedance Event is actually the same. In evaluating the risk of a pilot Exceedance Event in a certain flight fleet, we only need to calculate the probability of the Exceedance Event occurrence of the pilot, while evaluating the pilot operation performance for a period of time. It is possible to calculate the probability of the Exceedance Event occurrence of each parameter of the pilot in a period and compare it with the stable value of the corresponding Exceedance Event occurrence of the flight fleet. Finally, the pilot’s operation performance was evaluated by evaluating each of the pilot’s Exceedance Event risk levels. Based on the above analysis, taking the Boeing 737–800 model as an example, the evaluation model of pilot operation performance was written as follows:

$$p_{fleet,i} = \frac{D_{fleet,i}}{N_{fleet}} \quad (i = 100, 101, 102 \dots 207) \tag{1}$$

$$p_{fleet} = \frac{\sum_{i=100}^{207} D_{fleet,i}}{N_{fleet}} \quad (i = 100, 101, 102 \dots 207) \tag{2}$$

In formulas 1 and 2, $p_{fleet,i}$ is the probability of a Detect Limit event occurrence of the entire fleet. p_{fleet} is the probability of all kinds of Detect Limit events occurrence of the entire fleet. $D_{fleet,i}$ is the number of a Detect Limit event of the entire fleet. N_{fleet} is the number of the flights of the entire fleet. i is the code of QAR Exceedance Event, from 100 to 207.

$$p_{pilot,i} = \frac{D_{pilot,i}}{N_{pilot}} \quad (i = 100, 101, 102 \dots 207) \tag{3}$$

$$p_{pilot} = \frac{\sum_{i=100}^{207} D_{pilot,i}}{N_{pilot}} \quad (i = 100, 101, 102 \dots 207) \tag{4}$$

In formulas 3 and 4, $p_{pilot,i}$ is the probability of a Detect Limit event occurrence of a pilot. p_{pilot} is the probability of all kinds of Detect Limit events occurrence of a pilot. $D_{pilot,i}$ is the number of a Detect Limit event of a pilot. N_{pilot} is the number of the flights of a pilot. i is the code of QAR Exceedance Event, from 100 to 207.

$$P_{fleet,i} = \frac{A_{fleet,i}}{N_{fleet}} \quad (i = 100, 101, 102 \dots 207) \tag{5}$$

$$P_{fleet} = \frac{\sum_{i=100}^{207} A_{fleet,i}}{N_{fleet}} \quad (i = 100, 101, 102 \dots 207) \quad (6)$$

In formulas 5 and 6, $P_{fleet,i}$ is the probability of a Alert Limit event occurrence of the entire fleet. P_{fleet} is the probability of all kinds of Alert Limit events occurrence of the entire fleet. $A_{fleet,i}$ is the number of a Alert Limit event of the entire fleet. N_{fleet} is the number of the flights of the entire fleet. i is the code of QAR Exceedance Event, from 100 to 207.

$$P_{pilot,i} = \frac{A_{pilot,i}}{N_{pilot}} \quad (i = 100, 101, 102 \dots 207) \quad (7)$$

$$P_{pilot} = \frac{\sum_{i=100}^{207} A_{pilot,i}}{N_{pilot}} \quad (i = 100, 101, 102 \dots 207) \quad (8)$$

In formulas 7 and 8, $P_{pilot,i}$ is the probability of a Alert Limit event occurrence of a pilot. P_{pilot} is the probability of all kinds of Alert Limit events occurrence of a pilot. $A_{pilot,i}$ is the number of a Alert Limit event of a pilot. N_{pilot} is the number of the flights of a pilot. i is the code of QAR Exceedance Event, from 100 to 207.

3 System Design

In the last section, the flight operation performance evaluation model was established based on flight QAR data and quantitative evaluation method. There were 108 evaluation indexes, and we can select one or several of them for combination to evaluate the pilot’s flight operation performance. For example, it’s possible to use three landing operation performance evaluation indexes (touchdown distance, vertical acceleration, and pitch angle) [14, 15] of the pilot to evaluate the pilot’s landing operation performance objectively according to the model and algorithm.

In this section, the flight operation performance evaluation system will be introduced. The flight operation performance evaluation system was designed to include 7 modules: pilot management, QAR event inquiry and statistics, pilot operation performance evaluation, training and upgrade program, early warning, user center and system administration. The hierarchical structure of the system and each sub-function module of the system were shown in Fig. 2.

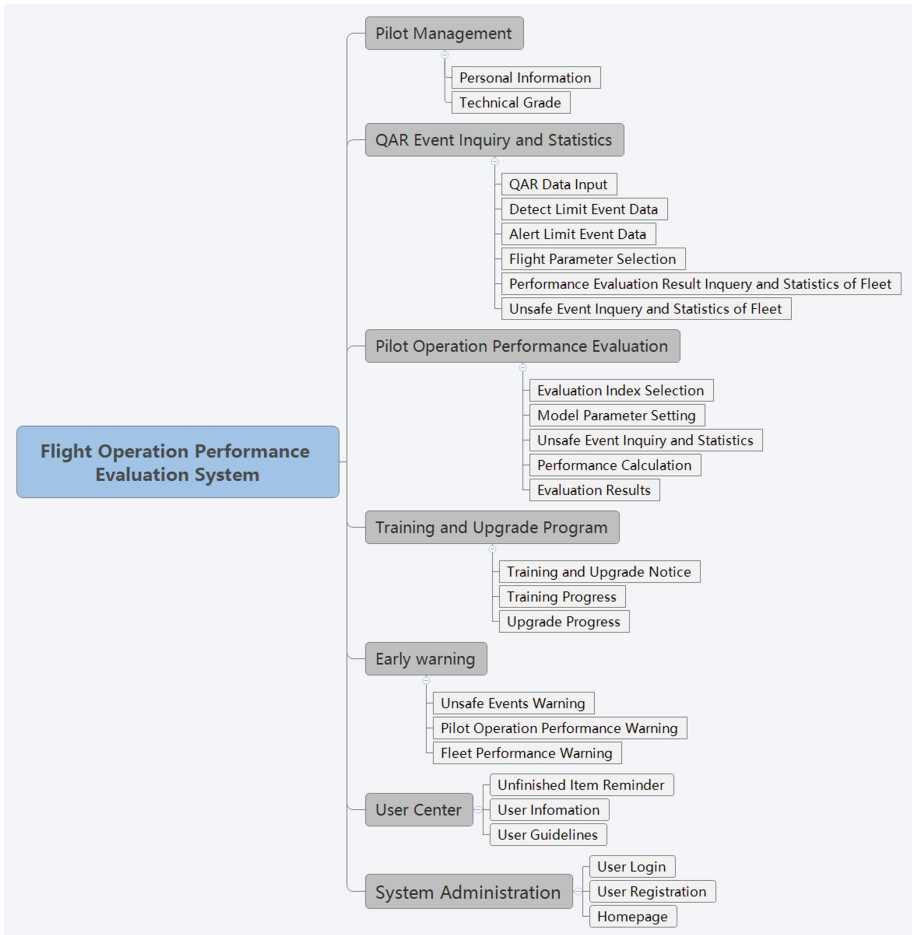


Fig. 2. Hierarchy diagram of flight operation performance evaluation system

4 System Development

4.1 Development Environment and Process

We adopt LNMP architecture, a popular web development technology to develop the database and system. The system is hosted by Linux operation system with Nginx as web server. MySQL is used as database server and we have one slave database hosted by another machine for data backup. We use PHP as our programming language. For the frontend development, we use bootstrap.css for both desktop and mobile friendly visiting.

4.2 System Interface and Functions

The developed Flight Operation Performance Evaluation System (FOPES) includes 7 modules, such as QAR event inquiry and statistics, pilot operation performance evaluation, training and upgrade program, and early warning. The main interface was shown in Fig. 3. The main interface includes a menu bar and links to 6 functional modules.

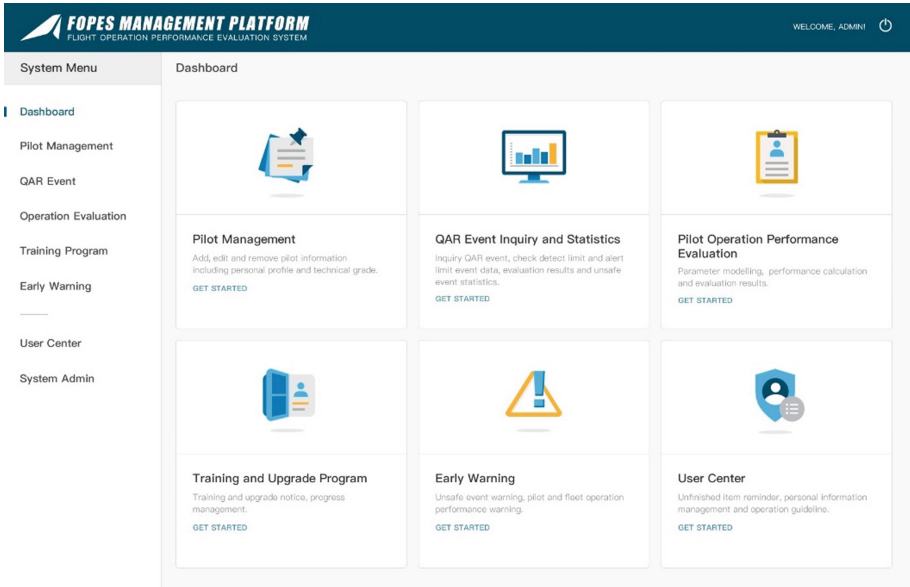


Fig. 3. Main interface of FOPES

The core function of this system is to evaluate the pilot's operation performance through using flight data and evaluation model. Firstly, clicking the "Pilot Operation Performance Evaluation" icon in the main interface. Next, set the parameters such as the pilot's name, the evaluation item and the time period. And then, the system will enter the calculation page. When the calculation is completed, the system will provide a prompt box and jump to the evaluation result page, as shown in Fig. 4. The evaluation results of this pilot can be sent to the training department for the targeted training.

Another important function of FOPES is to provide users with QAR event inquiry and statistics. Users can enter the event inquiry statistics page by clicking *QAR event inquiry and statistics* on the main interface. After inputting the information regarding the captain and the time period, the system will indicate the relevant statistical results, as shown in Fig. 5.

The user can not only inquire the relevant event information based on the entered information of captain, flight date, flight number, aircraft type, and event type, but also evaluate the pilot's operation performance by entering one or a group of parameters.

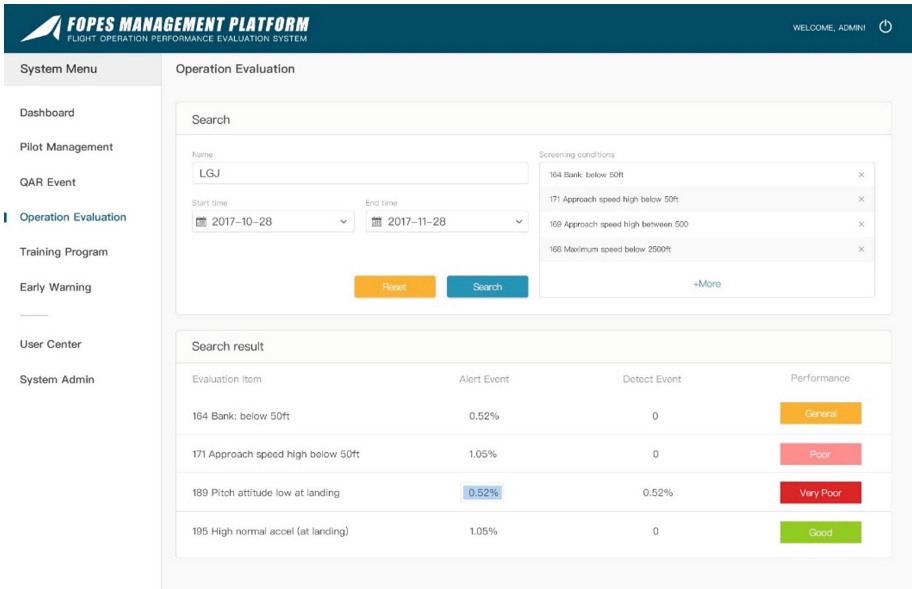


Fig. 4. Evaluation results of flight landing operation performance

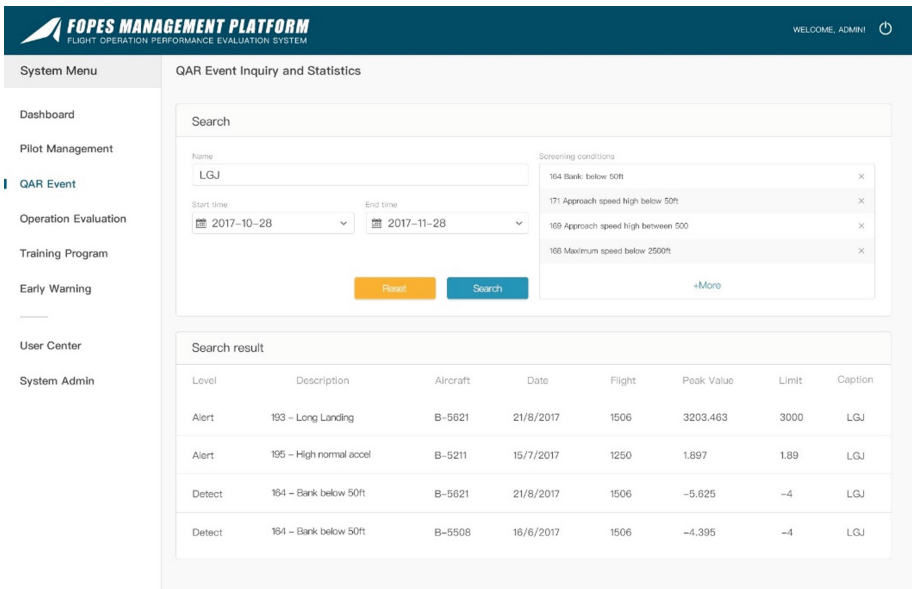


Fig. 5. One pilot’s QAR event inquiry and statistics

During the system daily operation, it will notify the user for subsequent processing if a warning occurs on the system.

5 Conclusions

The Flight Operation Performance Evaluation System was introduced in this study. The system has been tested in the flight quality control department of an airline, and it will be put into use. Flight performance evaluation experts set several evaluation indexes for combination to carry out the pilot's flight operation performance, recommendations for improvement, and other operations. The trial results showed the following:

- (1) The system can accomplish all basic functions, from the input of basic information and parameters to the output of evaluation results. It achieved QAR event inquiry and statistics, evaluation of the pilot's operation performance, training and upgrade program, warning and other functions earlier, indicating that the integrity of the system is good.
- (2) The system provides a support tool for flight operations quality monitoring and flight training management. The system can evaluate the pilot's operation performance from multiple dimensions, and that is more objective, effective, and reasonable. It will give an warning earlier and suggestions for improvement so that we can arrange follow-up training for the pilot. The airline's performance rewards and punishments would get a more accurate and objective basis from this system.
- (3) The system not only provides actual data support for flight operation department to monitor flight risk, but also provides effective basis and reference for flight training department to arrange targeted improvement training. However, the system needs to be improved for shortening its response time and processing. However, the system needs some improvement to connect with the other systems of the airline for data sharing. So that we can manage the pilot's operation performance more comprehensively and effectively.

References

1. Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., Wiegmann, D.: Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. *Hum. Factors* **49**(2), 227–242 (2007)
2. Ouraan, M.S., Shahin, B.N., Aqqad, S.S.: Human factors in Royal Jordanian Air Force five years experience. *Aviat. Space Environ.* **67**(7), 710 (1996)
3. Jarvis, S., Harris, D.: Development of bespoke human factors taxonomy for gliding accident analysis and its revelations about highly inexperienced UK glider pilots. *Ergonomics* **53**(2), 294–303 (2010)
4. Civil Aviation Administration of China: Annual Report of China Aviation Safety. CAAC, Beijing (2007)
5. Civil Aviation Administration of China: Annual Report of China Aviation Safety. CAAC, Beijing (2016)
6. Wickens, C.D., Hollands, J.G.: *Engineering Psychology and Human Performance*, 3rd edn. Prentice Hall Press, Upper Saddle River (2000)

7. Ruishan, S., Lei, W., Ling, Z.: Analysis of human factors integration aspects for aviation accidents and incidents. In: Harris, D. (ed.) EPCE 2007. LNCS (LNAI), vol. 4562, pp. 834–841. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73331-7_91
8. Civil Aviation Administration of China: Implementation and management of flight operation quality assurance. Advisory Circular: 121/135-FS-2012-45. CAAC, Beijing (2012)
9. Wang, L., Wu, C., Sun, R.: An analysis of flight quick access recorder (QAR) data and its applications in preventing landing incidents. *Reliab. Eng. Syst. Saf.* **127**, 85–96 (2014)
10. Wang, L., Ren, Y., Sun, H., Dong, C.: A landing operation performance evaluation system based on flight data. In: Harris, D. (ed.) International Conference on Engineering Psychology and Cognitive Ergonomics. LNCS, pp. 297–305. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58475-1_22
11. Wei, J., Zeng-Liang, L.G., Qi-Fu, P.B.: The method of quantitative area risk assessment and its application in chemical industrial park. *China Saf. Sci. J.* **19**(5), 140–146 (2009)
12. Wang, L., Wu, C., Sun, R., Cui, Z.: A quantitative evaluation model on hard landing risk based on flight QAR data. *China Saf. Sci. J.* **24**(3), 1–10 (2014)
13. Wang, L., Sun, R., Wu, C., Lu, Z., Cui, Z.: A flight QAR data based model for hard landing risk quantitative evaluation. *China Saf. Sci. J.* **24**(2), 88–92 (2014)
14. Wang, L., Wu, C., Sun, R., Cui, Z.: An analysis of hard landing incidents based on flight QAR data. In: Harris, D. (ed.) EPCE 2014. LNCS (LNAI), vol. 8532, pp. 398–406. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07515-0_40
15. Wang, L., Wu, C., Sun, R.: Pilot operating characteristics analysis of long landing based on flight QAR data. In: Harris, D. (ed.) EPCE 2013. LNCS (LNAI), vol. 8020, pp. 157–166. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39354-9_18



Pilot Performance Assessment in Simulators: Exploring Alternative Assessment Methods

Pete McCarthy¹(✉) and Arnar Agnarsson²

¹ Cranfield University, Bedford, UK
pete.mccarthy@cranfield.ac.uk

² Iceland Air, Reykjavik, Iceland
aja.crew@icelandair.is

Abstract. Flight crew performance and competency assessment are daunting tasks requiring expertise and training, and still will not be possible without a certain degree of subjectivity. On the other hand, collecting reliable data on flight crew competencies is at the core of Evidence-based Training, a major modernization of training methodology the industry as a whole has embarked on. Data from assessment informs training departments about where pilots seem to be lacking in proficiency, so those issues can be addressed in initial and recurrent training programmes of airlines. The effectiveness of the training hinges on the quality of the data. Accurate interpretation of the data is crucial for the decisions on training to respond to the true needs of commercial pilots. The industry has made a great effort to develop ways to measure crew performance. The role of Human Factors in incidents and accidents has been known for a long time, and the need to assess and train Human Factors has been identified. In recent years, with the introduction of Evidence Based Training (EBT) there has been a shift in focus from task-based assessment to competence based assessment. This study analysed crew performance in 25 videos from simulator sessions in a high fidelity full flight simulator. A checklist of Desired Flight Crew Performance (DFCP) was used to distinguish between high and low performing crews. Then the performance of selected crews was analysed in detail, using Performance Indicators (PI) as developed in EBT. The findings suggest that while the DFCP method was useful for the classification of high and low performing crews, the PI method provided detailed information for the understanding of underlying factors that affected the performance of the crews. The study also considers the value of using PI to understand and emulate well executed flying and problem solving, to change the focus of training from the study of error and accidents, to training best practices and safe operation.

Abbreviations

1. APK - Application of Procedure
2. CAA - Civil Aviation Authority
3. CAP - Civil Aviation Publication
4. CFIT - Controlled Flight Into Terrain
5. COM - Communication
6. CRM - Crew Resource Management
7. DFCP - Desired Flight Crew Performance
8. EASA - European Aviation Safety Agency

9. EBT - Evidence Based Training
10. FPA - Flight Path Management Automation
11. FPM - Flight Path Management Manual
12. FSTD - Flight Simulator Training Device
13. G/A - Go-Around
14. ICAO - International Civil Aviation Organisation
15. IRR - Inter-Rater Reliability
16. JAA - Joint Aviation Authority
17. KNO - Knowledge
18. KSA - Knowledge, Skill & Attitude
19. LOFT - Line Orientated Flight Training
20. LTW - Leadership and Teamwork
21. NOTECHS - Non-technical skills
22. PI - Performance Indicator
23. PIC - Pilot in Command
24. PSD - Problem Solving and Decision Making
25. SAW - Situational Awareness
26. SOP - Standard Operating Procedures
27. WLM - Workload Management
28. NTSB - National Transportation Safety Board

1 Introduction

In general, commercial air transport has become one of the safest modes of transportation in the last few decades. There have been advances in technology, which have reduced the accident rates to historically low levels. With this new, advanced and reliable technology, there have been some challenging issues regarding pilot training, and especially how pilots respond to unexpected events. In this study, a critical look is taken at two different methods of assessing commercial pilots in aviation. While traditional methods, based on observations performed by highly experienced and well trained flight instructors or examiners, can provide reliable data, the expertise of evaluators might vary in practice, which can cause variability of the results. This may become an issue for a data driven concept such as Evidence-based Training. What pilots do on the job is commonly expressed in technical skills, and human factors, which are commonly named non-technical skills. It can be challenging to assess pilot performance as it is difficult to isolate one from another. There are different assessment methods in place which are looked at in this paper, and there is empirical research which has been conducted that are considered as well.

The aviation industry has understood the importance of simulating the real-world experience through virtual environment. Simulation is used in many professions, and extensively for pilot training. For the training to be effective there is a need for a methodology that provides systematic and structured learning experiences. The effectiveness of this training is dependent on the quality of performance measurement practices in place. Performance measurement during each session must be diagnosed;

that is, the causes of effective and ineffective performance must be determined. This diagnostic measurement drives the systematic decisions concerning corrective feedback and remediation (Rosen et al. 2008). There are always challenges when it comes to assessing pilots during recurrent training. In Europe, every pilot must be assessed every 6 months based on EASA regulation. The regulation states that every pilot must be assessed in technical and non-technical skills.

2 Background

2.1 Simulator

In fast moving and complicating domains such as the military, medicine, business, and aviation, the workplace is characterized by high degrees of complexity and competitiveness. The need to maximize human performance is essential for safety, effectiveness, efficiency, and even survival. Human performance can be resilient, adaptive, and flexible in ambiguous and information-intensive contexts, but it can also be plagued with error and inefficiencies. Preparing people for performance in complex environments requires a complex approach to training (Salas et al. 2008). The simulator training in aviation has evolved somewhat throughout the last decades but in general, regulations that control the training have been very tenacious. In the early 1950's with the introduction of the jet age, there was a training scheme that was largely based on the evidence of hull loss from early generation jets so in a simple way the training developed into a system where every new accident triggered the introduction of a new task into already saturated training programmes in response to the accident. This could easily lead to an attitude where training was all about *ticking the boxes*, instead of responding to the real needs of pilots. Aircraft design and reliability has improved substantially over time, and at the same time the industry experienced accidents with hull losses where there was no malfunction of the onboard equipment. This gradually led to a shift in focus towards the human factor, or human failure. The human factor has been researched very extensively in the past three decades. According to researches, human failure has contributed to more than 2/3 of all accidents (Helmreich et al. 1999). A good example of that is the CFIT (Controlled Flight into Terrain), resulting in hull loss where inadequate situational awareness is almost always a contributing factor (ICAO 2013). In the late 1970's Crew Resource Management (CRM) started to be developed. The beginning of CRM is normally traced back to a workshop by NASA in 1979 (Helmreich et al. 1999). The CRM was intended to address Human Error. CRM was mainly classroom based, and later implemented into simulator programs like LOFT (Line Oriented Flight Training) etc. In the late 1990's the JAA precursor of EASA came up with Non-technical skill evaluation or NOTECHS to evaluate non-technical skills in simulator.

2.2 Simulator Assessment

Under EASA authority there is a requirement for technical and non-technical skills assessment of flight crew. EASA regulation stipulates the requirement for non-technical

assessment in AMC1 ORO.FC.115 Crew resource management (CRM) training “Assessment of CRM skills is the process of observing, recording, interpreting and debriefing crews and crew member’s performance using an accepted methodology in the context of the overall performance.” (EU 2012). EASA stipulates also that the method must be accepted by the national authority. Most operators are using the NOTECHS system or a similar system to fulfil this requirement. The NOTECHS system was issued in late 90’s and recommended by JAR.OPS (precursor of EASA). It was recognised that the task to assess non-technical skills was more subjective than assessing technical facts. The Notechs systems was intended to minimise that factor with using behavioural indicators to assist the raters (Flin et al. 2003).

There have been researches in the past regarding assessment in simulation based training. It is known in the military, medicine, business, and aviation world. In the commercial aviation world, there is an extra challenge, as the assessment is done both on the team and on individual performance. Aviation is a workplace characterized by a high degree of complexity and competitiveness. Therefore, maximizing human performance is essential for safety and effectiveness. The effectiveness is dependent on the quality of performance measurement practices in place. The effective and ineffective performance must be determined.

It is very important that performance is explicitly defined and measured. Otherwise it is impossible to change, or improve it systematically. This measurement gets more complicated with more complexity. Human performance is essentially behaviour in completing a task. The problem is to realise what behaviours are important components of performance (Salas et al. 2008).

The simulator assessment has evolved in the last two decades. In the beginning, NOTECHS was designed as a professional tool to be used by non-psychologists. It was not intended to judge flight crew personality or a toll for introducing psychological jargon (Flin et al. 2003). Examiners were not to fail pilots only on basis of non-technical skills, except associated with technical skills. This created some confusion and let the CRM be a kind of a stand-alone subject with its own requirements. The EBT competencies are a mixture of technical and non-technical skills and have their own performance indicators to assist examiners in assessing and debriefing with the intention of promoting learning. According to ICAO Doc 9995 there are 8 competencies and 59 performance indicators. One of the challenges is the methodology to assess pilots with 59 performance indicators. The data gathered from the system is the important bit as it gives a good picture for the operators to see where they should put in an effort, and how they should channel the resources in their training. But even the highly trained and most experienced examiners are limited in what they can reliably measure.

2.3 Evidence Based Training (EBT) Versus Traditional Training

There is a shift in training with the new method implemented by ICAO doc 9995 that introduces EBT training. This training is focusing on competency based training instead of task based training. It arose from an industry-wide consensus that reduction in aircraft hull loss and fatal accident is needed. It was necessary to review the existing recurrent and type rating training for commercial pilots. The EBT programme and

philosophy are intended as means of assessing and training key areas of flight crew performance in a recurrent training system (ICAO 2006).

Pilot core competencies were developed to support the Evidence-based Training (EBT) concept adopted by ICAO in 2013. An international industry working group was established in 2007 for the development of a competency-based approach to recurrent training and assessment. The first and critical step in the development of EBT was to identify a complete framework of performance indicators, in the form of observable actions or behaviours, usable and relevant across the complete spectrum of pilot training for commercial air transport operations. These competencies and performance indicators combine the technical and non-technical (CRM) knowledge, skills and attitudes that have been considered essential for pilots to operate aircraft safely, efficiently and effectively. A framework of behaviours was developed, divided into 8 core competencies, each with observable performance indicators. The competencies were published in the ICAO Doc 9995 Manual of Evidence-based Training. The core competencies are primarily an assessment tool, offering a different approach from the evaluation of outcomes and manoeuvres, the purpose being to understand and remediate root causes of performance difficulties, rather than addressing only the symptoms. The purpose of these performance indicators is to underpin the creation of performance expectations at all stages of training in a pilot's career. To complete the picture, a fair and usable system of grading performance is also required. The development of pilot core competencies was considered as the first important step towards the creation of the "total systems approach to training". By far the most significant challenges for operators using these competency frameworks is the creation of an effective performance assessment and grading system, and subsequently the need for instructor training and the assurance of inter-rater reliability (IATA 2014).

It is impossible to foresee all plausible accidents, especially in a complex and high reliability system, where the next accident may be something completely unexpected. Competency based training is trying to address this by shifting from pure scenario-based training, to prioritizing the development and assessment of key competencies. It uses the KSA (Knowledge, Skill, Attitude) to master the infinite number of competencies that allow a pilot to manage situations in flight that are unforeseen by the aviation industry. The EBT competencies encompass what has previously recognised as technical and non-technical knowledge, skills and attitude. The aim of EBT is to help pilots develop the identified competencies that are required to operate safely, effectively and efficiently in a commercial air transport environment (ICAO 2013).

2.4 Inter-Rater Reliability

A major challenge in conducting any kind of performance assessment is the development of strong inter-rater reliability, and consistency in the approach. It is of great importance for the system (ICAO 2013). Each scenario provides an opportunity for in-depth feedback for any of the 8 core competencies. In many such focused scenarios, a pilot is exposed to variations in his own working environment. Every pilot has opportunity to practise aspects of the same core competencies in different situations, which accelerates the acquisition of expertise in the complicated domain which aviation is. From an inter-rater reliability perspective, it is essential that what is measured is

based on reliable observation. Even highly trained and experienced observers are limited in what they can reliably rate (Rosen et al. 2008). It is therefore very important that the rater is trained in facilitation, debriefing, and in using the system to judge the technical and non-technical skills. It is important to understand what separates *exceptional* from *average* or *poor* performance among teams. So, it is important to understand how teams successfully interact to deliver superior performance. To evaluate superior performance, some evaluation tool needs to be in place. In the commercial aviation, there are normally two pilots who form the team. For the purpose of training the crew, it is necessary to diagnose differences between individual and team performance.

3 Research Question

It is critical for the aviation industry to have a valid assessment and measurement system in place for flight crew performance. For both individuals and the system, it is important to have valid and reliable assessment methods to reduce subjectivity to a minimum. In this paper, the purpose is to analyse flight crew performance based on two different methods.

Those methods are:

- Performance Indicators (PI) used in Evidence Based Training (EBT)
- Desired Flight Crew Performance (DFCP): a task based binary checklist

Is a binary Yes/No observation checklist usable to effectively reflect flight crew performance assessment?

Do the PI's reflect that specific scenario the same way as the DFCP method does, does it give more insight into what is the core of the problem versus indicating that task was done or absent?

What competencies helps high performing crews and hinder low performing crews in a challenging scenario which includes responding to the unexpected?

As the binary checklist if more straight forward and simpler to use, which recommendations can be made for improvement of the binary checklist approach with regards to use in training?

4 Methods

4.1 Aim

The aim of the experiment was to create a scenario which includes elements which are normally not covered in routine training programs and would enable the researcher to monitor pilot performance in unexpected events. The experiment took place in a full flight, approved level D simulator, which enabled the crew to react exactly as they are used to in traditional training.

The researcher was provided with 25 exclusive access research videos of flight simulator events, by a large airline with a multicultural pilot workforce. Each crew was

exposed to the same scenario. The scenarios offered considerable challenges with elements of surprise and operational complexity. The presented videos were extracted from a real simulator environment. The pilots showed authentic nonscripted behaviours as they were not able to prepare themselves beforehand, but had to make fast decisions. Each video is about 30 min long for the whole scenario. Five of the videos were discarded as the content was obscured or sound corrupted which hindered the researcher to get a clear understanding of what was going on and made thorough analysis impossible.

4.2 Scenario

The scenario started at descent into an international airport in the Middle East. Three key events in the scenario formed the unexpected elements which were studied. In the first event, the crew (Captain and First Officer) were given clearance for an instrument approach. During the approach, the tailwind exceeded the limit for the aircraft, and this should instigate the crew to do a Go-Around; or later, for the crews that proceed with the approach, a loss of visibility below decision height forcing them to Go-around. During the Go-Around, the second event occurred, which was a subtle heading control failure in the automation, forcing manual reversion to regain heading control. During turn the final event took place: a bird strike hitting both engines, causing internal damage. The damaged engines would surge and stall until thrust was reduced on both engines. The crew were free to select any appropriate response to the failures. The pilots were faced with automation limitations and partial engine power on both engines. The six flight phases were the following:

- Instrument Approach (first approach)
- Go-Around
- Heading failure in automated system (subtle)
- Bird strike affecting both engines
- Planning the subsequent landing
- Approach and landing.

4.3 Analysis

The study was carried out in two steps: First an assessment of the performance of all crews based on a DFCP checklist; and then in step 2, the scores of the DFCP assessment were used to select the three highest, and three lowest performing crews, which were analysed further with Performance Indicators.

Step 1: Initially an objective method was selected to reduce or eliminate subjectivity. It is a known procedure to objectively measure the performance of pilots while undertaking certain tasks, like flight path management, or procedure adherence. It is more challenging to objectively measure the non-technical skill parts, like problem solving or situational awareness. To assist in that a known method was used: Desirable Flight Crew Performance (DFCP). This method is used to determine the action of the pilots against a safe outcome in the scenario. The DFCP list uses ranking of the flight crew against expected behaviour in the scenario. The industry is known for thorough

operating procedures or guidelines and training which were considered (Field et al. 2016). A list of 25 items was generated for this specific scenario. To increase validity in the choice of items on the list of tasks, two experienced Airbus simulator instructors were consulted in addition to the author who is experienced simulator instructor as well. The list was made up independently at first and then compared and finalised. As the videos were exclusively for the author, because of confidentiality issues, the other experts could only theoretically assume the tasks needed for a safe outcome. They were given details of the scenario and used their experience to suggest the tasks that were considered important. There was agreement about all the items in the list. The DFPC list is divided into six phases see Fig. 1.

| Phase | No. | DFCP: |
|----------------------|-----|---|
| First approach | 1 | Brief low visibility approach |
| First approach | 2 | Approach checklist |
| First approach | 3 | Arm Approach |
| First approach | 4 | Landing checklist before 1000’. |
| First approach | 5 | Clearly verbalize tailwind |
| Go-Around | 6 | GA due to tailwind (Critical) |
| Go-Around | 7 | GA due to loss of visual contact |
| Go-Around | 8 | Execute G/A actions - G/A Procedure (Flaps&Gear) |
| Heading failure | 9 | Verbalize heading failure |
| Heading failure | 10 | Execute immediate manual flying and turn right |
| Birdstrike | 11 | Verbalize birdstrike |
| Birdstrike | 12 | Verbalize surge/stall or ECAM actions |
| Birdstrike | 13 | Reduce power to try and stop surge |
| Birdstrike | 14 | Run correct Abnormal checklist (ECAM), ENG STALL |
| Birdstrike | 15 | Mayday call |
| Planning | 16 | Inform cabin crew of birdstrike/immediate landing |
| Planning | 17 | Review of aircraft/system status, options |
| Planning | 18 | Get info from ATC on WX, RWY avail |
| Planning | 19 | Assessment of RWY for second landing based on WX |
| Planning | 20 | No unnecessary delay for second approach |
| Planning | 21 | No landing with tailwind |
| Approach and landing | 22 | Verbalize energy management considerations |
| Approach and landing | 23 | Briefing, G/A no option |
| Approach and landing | 24 | Execute landing checklist |
| Approach and landing | 25 | Landing with flaps 3 |

Fig. 1. A list of 25 items divided into six phases in the scenario.

Each item on the list is considered essential for a safe outcome. The researcher watched all the videos and kept a score about items that were either completed or not completed. It was a score against the DFPC list, which included a number of safety critical items already identified. The list was used to collect each item and thereby draw a comparison between the crews. The rating was simply the sum of the DFPC items that were performed by relevant crew. The result was used to identify the high performing crew and low performing crew. Out of the twenty crews, one crew opted to land after the first approach despite having no visibility. For that reason, the crew was not included in the final comparison.

Each crew was scored according to the desired parameters, and total score calculated to select the three best performing crew and three least performing crew. One point was given for completing the task, and no point if the task was absent. For two of the items in the DFCP list “Go-Around due to tailwind” and “Reduce power to try and stop the surge” were considered critical and were given the weight of three points due to their role in the safe outcome of the exercise. Maximum score for each crew was therefore 28 (the crew got 3 points for *Go-around because of tailwind* and in case of *Go-around because of no visibility at decision height* one point was given).

The DFCP analysis only shows if certain task was completed, or absent. It does not explain why certain crews performed better than others according to the DFCP list.

Step 2: The second phase of the analysis tries to determine the differences in behaviour using the performance indicators (PI’s) that are used in EBT for behavioural measurement. As previously mentioned, the PI’s form observable behaviours and have been established by ICAO. Some modification is recommended for operators to adjust the PI to their specific needs in evaluation or assessments (Iata 2013). The PI’s used in this experiment were adjusted by EBT Foundation and LOSA collaborative in a joint venture, and one competency added to the 8 established competencies. The added competency is *Knowledge* and it has its PI’s as well. List of all competencies and their PI’s can be found in appendix. The PI’s are intended as a guiding tool to look beyond the outcome and subsequently look at the process that either hindered or helped the applicable crew in that specific scenario.

The researcher took the three-highest performing crew, and three lowest performing crew according to the DFCP and compared them. They were analysed by using observation and tangle them with applicable PI. Each PI was either positive (P) or negative (N) for each segment of the scenario. As the PI’s are globally designed to capture as much as possible in the real world, some of the PI were not relevant to this specific scenario and therefore not observed. There are total 66 performance indicators that were consulted for each six phases of each scenario. The observation was made against the individual and not as the crew as a collective unit. This method is considerably different from real environment as the researcher was able to use pauses, rewind etc. extensively. This allows the analysis to be very detailed. The PI’s were used as a checklist in effort to reduce the subjectivity because of one rater.

Each phase was then compared between the crews to try to establish what competencies were hindering, or helping the crew to deal with the situation.

5 Results

During this challenging scenario, all crew except one managed to land the aircraft on the runway. It was very different between crews how they handled the situation and situational awareness at first glance seemed to be lacking for low performing crews. Each crew had to spend some time on analysing what was actually going on, and this provided an opportunity to notice some differences. As all the crew were doing this on voluntary basis, no crew actually had engine failure on both engines, which would have been a more dramatic ending. That could however be expected in an actual scenario if

the crew does not reduce the engine power below the engine EGT (exhaust gas temperature) limit. After analysing the three-high performing crew and three low performing crew, a clearer better picture was starting to form. The intension was to compare the DFCP list with the Performance Indicators and see if there would be the same outcome.

A descriptive analysis was performed. The results of the two methods indicated a similar outcome. Both indicated that low-performing crew and high-performing crew scores were concurrent. The DFCP method is a checklist composed of items that were considered necessary for a safe outcome. There were considerable differences in the performance of the crews. After all the videos had been watched, a score based on the DFCP list was generated. See Fig. 2.

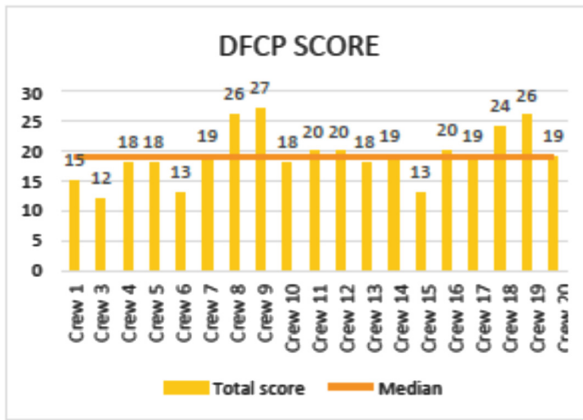


Fig. 2. DFCP total score for all 19 crews analysed. Median line for reference

The analysis gave a strong indication of differences in performance, see Fig. 3. The range of scores were from 12 as the lowest, to 27 the highest. Average is 19.15 and median value was 19. This indicates that there were considerable differences between high and low performing crews, and fairly high spectrum between high and low. High performing crews scored on many of the tasks required, or almost all of them. Low performance crews scored much less where tasks were commonly absent. It turned out that the decision to give two tasks in the DFCP list more weight than others, based on their criticality, did not affect the selection of crews. DFCP could distinguish between high and low performance crews but could not show reasons for crew actions. This will be further discussed in discussion and conclusion.

Figure 4 Drafts up simplified but typical track of the scenario. The green dotted line demonstrates track chosen for typically higher performance crews while the lower performance crews typically followed the red dotted line after the go-around. The pattern indicated by the red dotted line meant that the crew was landing with more tailwind then the aircraft is certified to do, while the green dotted line demonstrates the preferred track. This figure is just for demonstration it does not depict the exact track nor all courses of action decided by each crew.

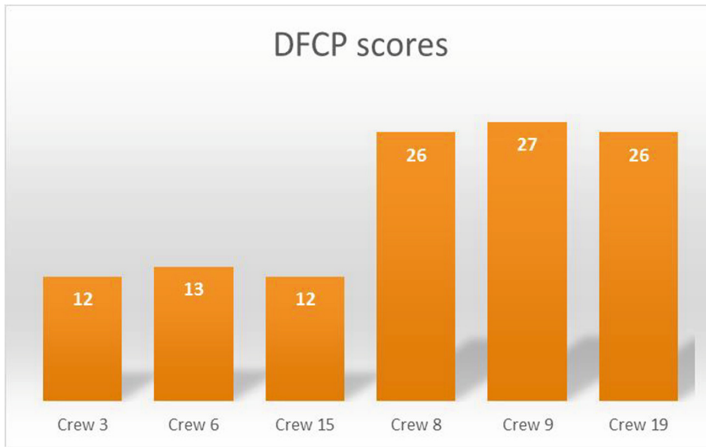


Fig. 3. Score for three highest performance crews and three lowest performance crews.

The result of the detailed PI analyses is presented, based on the numbers of (P) positive and (N) negative scores given in the evaluation. The score is made against observation that tangles with PI's and is either positive or negative. This allowed to obtain an overall score for each flight segment. Some of the PI are not observed. Each flight phase is scored separately in all competencies to see the difference when crew is faced with unexpected scenario, like bird strike and planning. Figures 5, 6, 7, 8, 9 and 10 show the difference in observed PI's in each flight phase.

The results show that in the first two phases there were not much differences between observed crew PI's. There are negative indications which are related to not detecting the wind change on the approach, which affected the situational awareness and decision making. When crews are faced with standard operation or threat they can expect (typical training scenario which pilots are frequently exposed to in simulator training e.g. Approach and Go-arounds etc.) they seem to handle that within certain criteria. Not much difference is detected between high and low performing crew. On the other hand, when unexpected events are introduced, there is more difference detected. Where the heading failure is introduced, which is subtle failure in the automation, a greater difference is detected. One crew was hesitant to revert to manual flying and therefore scored negatively. Some difference is detected there but not as much as would be expected based on the DFCP. During bird strike and subsequent planning phase, considerable difference is detected in PI's. High performing crew demonstrates more positive PI's but there is some evidence of negative PI's. That would be expected as the scenario is challenging and not typical for pilot routine training. The low performing crew is scoring high on negative PI's during these phases and reduction in positive scores are detected.

To explore further and see which competencies were helping or hindering the crew, a comparison was also made regarding all the 9 competencies. In Figs. 11, 12, 13 and 14 which is presented as a heat map to visually demonstrate the mostly affected competencies.

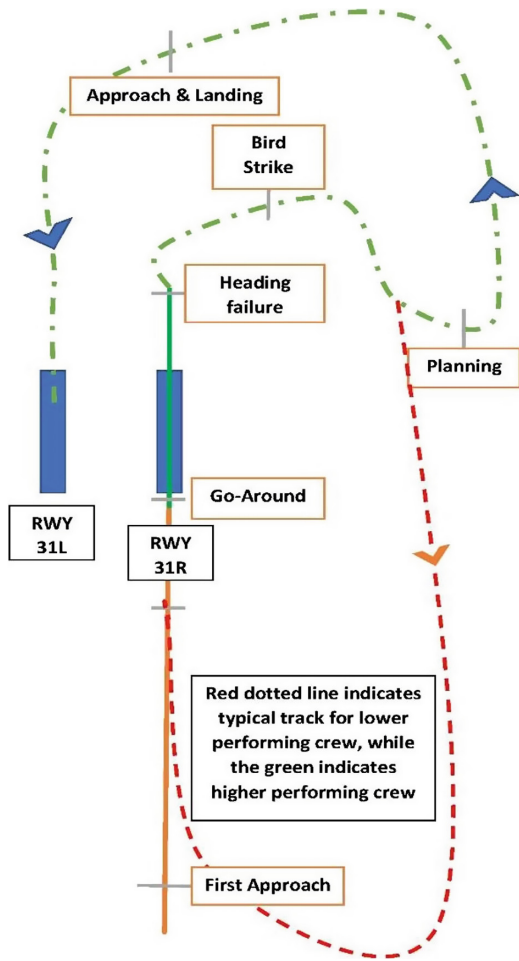


Fig. 4. Two typical tracks decided by crews (Color figure online)

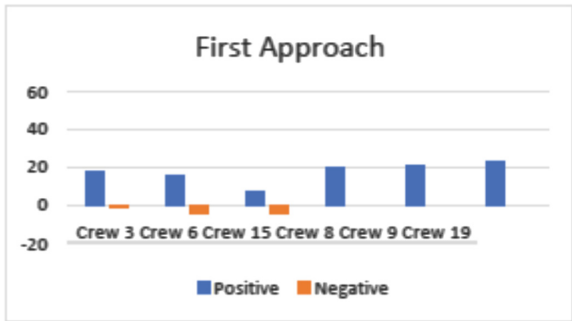


Fig. 5. Performance indicators identified for First Approach

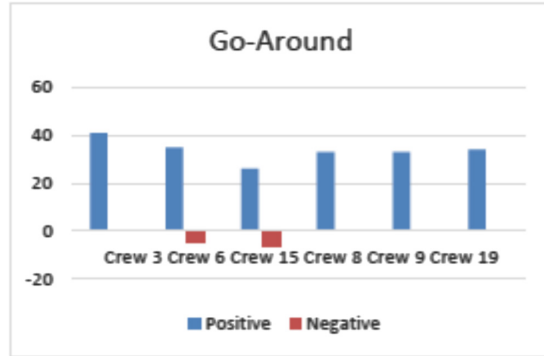


Fig. 6. Performance indicators identified for Go-Around

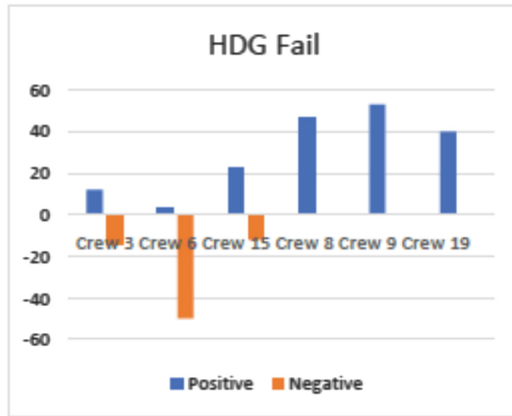


Fig. 7. Performance indicators identified for Heading Failure

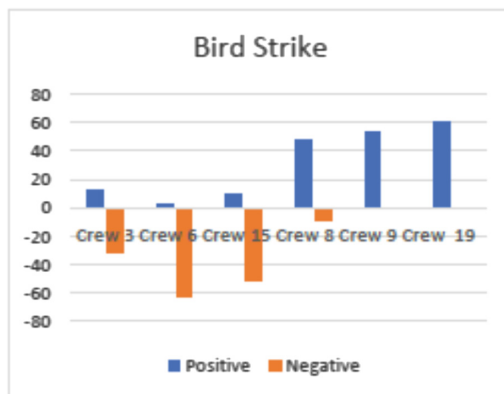


Fig. 8. Performance indicators identified for Bird Strike

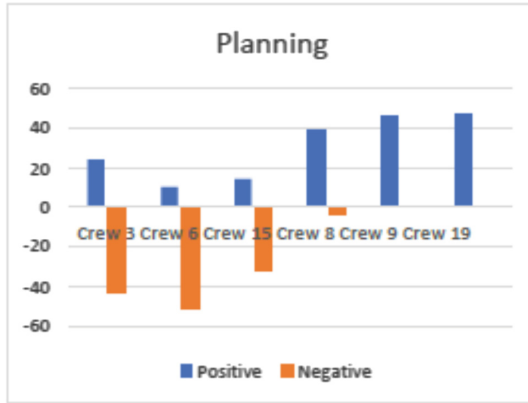


Fig. 9. Performance indicators identified for Planning

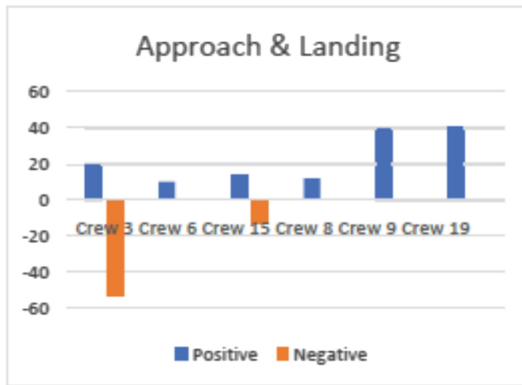


Fig. 10. Performance indicators identified for Approach & Landing

| Positive - High Performance Crew | | | | | | |
|----------------------------------|----------------|-----------|-----------------|-------------|----------|--------------------|
| Phase | Short Approach | Go-Around | Heading Failure | Bad traffic | Planning | Approach & Landing |
| Competence | | | | | | |
| APK | 18 | 18 | 16 | 18 | 14 | 17 |
| COM | 30 | 31 | 28 | 27 | 26 | 16 |
| FPA | 0 | 10 | 3 | 2 | 0 | 0 |
| FPM | 0 | 1 | 13 | 11 | 10 | 6 |
| KNO | 4 | 13 | 14 | 13 | 16 | 8 |
| LTW | 0 | 8 | 15 | 28 | 20 | 15 |
| PSD | 0 | 1 | 19 | 23 | 14 | 8 |
| SAW | 12 | 12 | 12 | 15 | 12 | 11 |
| WLM | 0 | 6 | 20 | 25 | 20 | 12 |

Fig. 11. Demonstrates the positive performance indicators that were analysed with the three high performing crews

| Positive - Low Performing Crew | | | | | | |
|--------------------------------|----------------|-----------|--------------------|--------------|----------|--------------------|
| Phase | Short Approach | Go-Around | Finalising Actions | Final Update | Planning | Approach & Landing |
| Competence | | | | | | |
| APK | 14 | 16 | 8 | 2 | 3 | 7 |
| COM | 21 | 24 | 7 | 3 | 10 | 8 |
| FPA | 0 | 15 | 6 | 0 | 0 | 2 |
| FPM | 0 | 2 | 5 | 6 | 9 | 9 |
| KNO | 0 | 10 | 2 | 1 | 4 | 2 |
| LTW | 0 | 10 | 6 | 3 | 9 | 8 |
| PSD | 0 | 3 | 0 | 6 | 4 | 0 |
| SAW | 6 | 12 | 3 | 1 | 5 | 6 |
| WLM | 0 | 10 | 1 | 2 | 4 | 2 |

Fig. 12. Demonstrates the positive performance indicators that were analysed with the three low performing crews.

| Negative - High Performance Crew | | | | | | |
|----------------------------------|----------------|-----------|--------------------|--------------|----------|--------------------|
| Phase | Short Approach | Go-Around | Finalising Actions | Final Update | Planning | Approach & Landing |
| Competence | | | | | | |
| APK | 0 | 0 | 0 | 0 | 0 | 0 |
| COM | 0 | 0 | 0 | 0 | 2 | 2 |
| FPA | 0 | 0 | 0 | 1 | 0 | 0 |
| FPM | 0 | 0 | 0 | 0 | 0 | 0 |
| KNO | 0 | 0 | 0 | 4 | 2 | 0 |
| LTW | 0 | 0 | 0 | 0 | 0 | 0 |
| PSD | 0 | 0 | 0 | 4 | 2 | 0 |
| SAW | 0 | 0 | 0 | 0 | 0 | 0 |
| WLM | 0 | 0 | 0 | 0 | 0 | 0 |

Fig. 13. Demonstrates the negative performance indicators that were analysed with the three high performing crews

| Negative - Low Performing Crew | | | | | | |
|--------------------------------|----------------|-----------|--------------------|--------------|----------|--------------------|
| Phase | Short Approach | Go-Around | Finalising Actions | Final Update | Planning | Approach & Landing |
| Competence | | | | | | |
| APK | 4 | 2 | 6 | 11 | 11 | 5 |
| COM | 0 | 2 | 20 | 29 | 23 | 15 |
| FPA | 0 | 3 | 9 | 0 | 0 | 2 |
| FPM | 0 | 0 | 5 | 4 | 9 | 2 |
| KNO | 0 | 0 | 4 | 16 | 14 | 7 |
| LTW | 0 | 2 | 7 | 19 | 11 | 10 |
| PSD | 0 | 2 | 5 | 24 | 24 | 11 |
| SAW | 6 | 0 | 9 | 13 | 12 | 8 |
| WLM | 0 | 0 | 11 | 28 | 22 | 13 |

Fig. 14. Demonstrates the negative performance indicators that were analysed with the three low performing crews

The heat map indicates that high-performing crew were scored positively in Communication (COM), Leadership and Teamwork (LTW), Problem Solving and Decision Making (PSD) and Workload Management (WLM).

Indication for the low performing crew is similar for positive score in the first two phases or slightly lower. Negative indication is very low for the first two phases which indicates that the low performing crews are performing normally under familiar, or expected events. This indicates that it is hard to detect the difference between the high and low performing crew under expected or familiar events. When unexpected events come into play, the positive scoring drops rapidly for the low performing crew. The outcome of the negative scores supports the picture as well: Low performing crews score negatively especially with unexpected events, which is reflected in Communication (COM), Leadership and Teamwork (LTW), Problem Solving and Decision Making (PSD), and Workload Management (WLM). This is quite opposite of the results of the high performing crews. The four identified competencies are helping the high performing crews, while at the same time hindering the low performing crews, indicating that certain non-technical skills are required to deal with this unexpected event. Although the high performing crews are scoring much more favourably, there is an indication of minor rise in negative PI score during a challenging situation, which would have been expected.

6 Discussion

The DFCP method was used to rate the performance of the crew in safety related decisions and actions. There was consensus from three experts regarding each task on the list. The list of tasks was successfully used to identify the desirable and less desirable actions taken by flight crew in response to expected and unexpected events they encountered in the simulator. DFCP could distinguish well between high-, and low performing crews that were subsequently analysed further with the Performance Indicators. The DFCP identified the flight crews that performed well but it did not explain potential reasons for the decisions, actions or differences in performance. To investigate these descriptions of performance and to investigate the potential reasons behind it an additional analysis was performed.

Reflecting on the results from the performance indicator analysis, low performing crew struggled in the competencies where high performing crews were strong. This was mostly evident in high workload situations. It was evident in the planning phase that low performing crew were suffering from inadequate communication which resulted in poor planning and lack of situational awareness. From the heat map (Figs. 11–14) where each competency is presented, Leadership and Teamwork, Problem Solving and Decision Making, Workload Management, and Communication are the ones that are most prominent. The competencies cannot be viewed as separate, independent entities. They are interconnected, and some of them are a consequence of good or bad in others. Communication is a good example: Communication is not an indicative of good or bad performance in itself, rather it propagates positive or negative behaviour in other competencies.

Communication was integrated into the four categories of non-technical skills in the NOTECHS, and not used as a separate competency (Flin et al. 2003), which, based on the findings of this study, might have left out some valuable data. Studies have found that high performing crews discussed in-flight problems more thoroughly than low performing crews. They also used low workload phases to plan ahead and anticipate by using strategy and planning. They were found to use fewer commands during a high workload situations (Grossman and Salas 2011).

As all the crews in this study are working for the same operator, and are current and active pilots on the type. This means that they are all trained and used to the same operating procedures. There was no clear difference between the crews when dealing with expected events. Contradicting to the other study, the low performance crews were not obviously detected in the first two phases. This concludes a slightly different picture as that had been detected before from other study on similar subjects, that the competencies did not show that much of a difference during routine operation between high and low performance crews (Field et al. 2016).

Events and tasks in the first two phases of the scenario are what pilots are exposed to in their regular recurrent training. This supports that training should focus on unexpected events, frequently exposing pilots to demanding situations, requiring them to utilise all the spectrum of the core competencies. The collection of observed competencies makes it possible to draw a clear picture of the difference between high and low performing crews.

The DFCP is essentially a checklist of tasks performed or not performed. Once the list of tasks has been agreed on, the marking is relatively straightforward, and different observers are likely to check the same boxes so there is a high likelihood of consistency in marking. It was more relevant to use that method as the use of rewind and pauses was not used extensively, which lowered the work load on the researcher. In testing terminology: the method is reliable. Performance indicators are observable behaviour which reflect the competencies as described in EBT. As the goal of the PI's is to reveal root causes to performance below, at, or above the expected level, this method probes deeper than DFCP. It seeks to give the observer an understanding of why a task was completed or omitted, not just if. However, as there is substantial judgement involved on behalf of the instructor, and a high probability of rater bias, the instructors need to be trained and standardised. Again, in testing terminology: this method presents challenges in reliability, but it scores high in validity because it attempts to capture things that are important to assess, not just those that are easy to assess. However, during this study the author could use the checklist in a different way than would be possible in simulator training, as the use of pauses, rewind etc. were useful. Even if the validity composes some real challenges, the positive thing is getting to the root cause. It would have been likely that at first glance the Situation Awareness would have been the cause of some of the failures. Situation Awareness is one of the labels in CRM which is commonly referred to. This label is commonly referred to in accident investigations as well, where the probable cause was loss of situational awareness. The performance indicators in this scenario, although detecting some negative performance in situation awareness, indicate that the crew was lacking in other domains more dominantly. These evidences indicate that it is important to retrain and debrief the crew to help with transfer of training. Behind the terms, human error and situational awareness is another

psychological world to do with attention, perception, decision making, and so forth. Human factors have produced or borrowed terms that try to capture these phenomena. Complacency, situation awareness, crew resource management, shared mental models and workload are common currency today that are deep rooted in science, but at the blunt end, people have difficulties putting finger on, and don't dare to ask what they actually mean (Dekker 2006). The importance of the performance indicators is training the instructors. They are not intended to apply a psychological jargon; the performance indicators are simply a behavioural marker that are intended to tangle with observations that are made by the instructor or examiner. From this study, there are detailed data which would not be expected in normal recurrent simulator, where the simulator instructor is not only observing the crew. The simulator instructor is also occupied running the simulator, acting as Air Traffic Controller (ATC), playing cabin crew, and so forth. For the instructor, it is important that he acts like he normally does, observes and records what he sees from the crew, and after the simulator, he takes the recording and observation and applies it to the performance indicators. This assists the instructor by not overloading him with tasks and the performance indicators are something which he should be able to observe. As mentioned before, the instructor training is very important as the data gathered in the simulator is only as good as the validity of the data. For training transfer of the pilots, which might add is the ultimate goal, facilitation debriefing technique is considered essential for training cognitive ability and training transfer (Grossman and Salas 2011). It is important for the student or pilot to understand his weaknesses or the root cause instead of simply stating the label of non-technical skill that needs improvement. Same apply to high performing crew, realising why things went well is critical to motivation and transfer of training (McDonnell et al. 1997).

As explained above, two methods were used to analyse crew performance in this study: DFCP (Desired Flight Crew Performance) checklist; and Performance Indicators. One of the videos revealed an interesting aspect of the different things that the two methods capture.

The above-mentioned video was very interesting with respect to the two methods. The crew performed nearly all the DFCP tasks – but crashed the aircraft. The list did not specify that the aircraft had to land on the runway. This crew scored above average, but the outcome was unsafe, to say the least. So, the conclusion in this scenario, the method is lacking in validity. The PI was much more effective in capturing deficiencies in the performance of this crew. Although the crew did many things well, they would have gotten a negative indication for Application of Procedure, Problem Solving, Situation Awareness, and Communication.

It is outside the scope of this paper, but it would be very interesting to try to improve the DFCP list and link the items to performance indicators. The DFCP list served its purpose to distinguish between high and low performance crews but to combine these two methods would assist the instructor. If successful, that would strengthen the overall validity and reliability of the combined method and enable the instructor to capture more information. This was given some consideration but there seem to be some difficulties in doing that, and it might be impracticable. For example, the PI's are based on individual performance, but the DFCP assesses tasks performed by the crew. This would need to be solved.

One of the competencies that is identified strongly in this experiment was Leadership and Teamwork (LTW). The Teamwork competency explains the effective teamwork which has lately attracted more attention and interest in researches. Empirical evidence suggests, that there is an increase in requirement for teamwork in complex domains. Teamwork studies have indicated that individuals become less willing to accept input or feedback from their team members when in high workload situations. Teams that are oriented have shown to perform better under high pressure or high workload. Team orientation is based on effective communication, decision making and being able to manage workload (Salas et al. 2008). Although there is not full consensus on Teamwork researches, this coincides with the data that were received in this study on identified competencies where the high performing crew scored positively in these domains and where the low performing crew scored negatively. We are seeing this study supporting this empirical finding. With closer look, there is also another empirical finding where the study can be connected: Safety II is a recent term which describes safety from another point of view. Safety II is defined as the ability to succeed under both expected and unexpected conditions. It becomes a characteristic of how systems function and is based on that the human is clearly an asset rather than a liability (Hollnagel 2014). With the Performance Indicators, more accurate data will be received to study both negative and positive performance. With the positive data, there is a possibility to study expertise, taking into account which competencies are dominant in different scenarios, and use that data to look more closely at safety II. This way it could help to achieve more total system approach to training. It has the potential to be beneficial for operators that are willing to adopt this way of thinking as is defined in Safety II. Current safety systems for a typical airline today are relying on flight data that it receives from their own aircraft (Flight Data Monitoring). These data are in fact quite similar to the DFCP. They are real data and can tell you what went on, but they are limited as a tool to explain the reason. If it is possible to study why things go right, the data given from the performance indicators in EBT will be very valuable. In this study, the analysis done with the performance indicators give a detailed picture of the performance, hence they can be used to create a feedback loop into the training and safety system. They are currently being used for investigating human (BEA 2013). What is further possible is to use the data to study expertise per se, so we know why things are going right and continue from there.

7 Limitations

The author is experienced simulator examiner and instructor. Nevertheless, is the research analysis very different from real simulator environment, the ability to use rewind and pauses to thoroughly look at whatever is said and tone of voice gives more detail in analyses. It was new for the author to be able to look at body language as the details are very subtle. For example, hand gesture, etc. It also possible that confirmation bias took place after the selection of low and high performing crew had taken place. Knowing that might have influenced the result of performance indicators. In order to reduce or hopefully remove the confirmation bias, the discussions among the crew were mostly written up to get better understanding and being aware that things might be biased.

8 Recommendations

Further studies on training transfer is needed. Further knowledge is needed to study if there is a transfer in KSA (Knowledge, Skill, Attitude) between different scenarios or domain. There are positive indication that training transfer is likely specially in cognitive skills, so the cognitive ability is transferring to other domains (Grossman and Salas 2011).

9 Conclusions

Systematic gathering of valid data about crew performance is essential for effective training and safe flight operations. A binary yes/no checklist is usable to assess flight crew performance, but it has serious limitations as, while it gives useful data about the completion of tasks, it is not a suitable tool to analyse why a task is not completed. As the checklist is used in this study it may fail to capture important information so it needs to be developed further.

The two methods that were used in this study were similar in their ability to detect high-, and low performing crews. However, the information they provide the researcher with are very different. The Performance Indicators were very effective in identifying the underlying competencies that helped, or hindered the crews. This can help training departments to make more informed decisions about training needs.

The competencies that help high performing crews are, LTW, COM, PSD, and WLM. Conversely, lack of those competencies was a barrier to low performing crews.

The use of Performance Indicators poses challenges in rater reliability because instructor bias is likely to cause subjectivity. These challenges can be mitigated with effective training of instructors. The performance indicators are very useful to capture the competencies, which are important for pilots to have to effectively master complicated, unexpected situations. The study detected more negative performance indicators in competencies that are more related to non-technical skills rather than technical skills for the low performing crews. The difference between high and low performing crews in technical skills was not as evident, which indicates that training needs to be more effective in non-technical skills, despite the effort that the industry has made e.g. with the emphasis on NOTECHS and CRM training in the last few decades. Conclusions should state concisely the most important propositions of the paper as well as the author's views of the practical implications of the results.

References

- Barry Issenberg, S., MCGAGHIE, W.C., PETRUSA, E.R., LEE GORDON, D., SCALESE, R.J.: Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med. Teach.* **27**(1), 10–28 (2005). <https://doi.org/10.1080/01421590500046924>
BEA: Sx-Bhs, August 2013

- Cannon-Bowers, J.A., Salas, E.: Team performance and training in complex environments: recent findings from applied research. *Curr. Dir. Psychol. Sci.* **7**(3), 83–87 (1998). <https://doi.org/10.1111/1467-8721.ep10773005>
- Dekker, S.: The Field Guide to Understanding Human Error. *Ergonomics*, vol. 51 (2006). <https://doi.org/10.1080/00140130701680544>
- Ericsson, K.A., Ward, P.: Capturing the naturally occurring superior performance of experts in the laboratory: toward a science of expert and exceptional performance. *Curr. Dir. Psychol. Sci.* **16**(6), 346–350 (2007). <https://doi.org/10.1111/j.1467-8721.2007.00533.x>
- EU: Commission Regulation (EU) No. 965/2012. Official Journal of the European Union, 5 October 2012
- Field, J.N., Mohrmann, F., Fucke, L., Grácio, B.C.: Flight crew response to unexpected events: a simulator experiment. In: *AIAA Modeling and Simulation Technologies Conference* (2016). <https://doi.org/10.2514/6.2016-3373>
- Flin, R., Martin, L., Goeters, K.-M., Hörmann, H.-J., Amalberti, R., Valot, C., Nijhuis, H.: Development of the NOTECRS (non-technical skills) system for assessing pilots' CRM skills. *Hum. Factors Aerosp. Saf.* **3**(2), 95–117 (2003). http://www.safetylit.org/citations/index.php?fuseaction=citations.viewdetails&citationIds%5B%5D=citjournalarticle_37801_6, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.472.6866&rep=rep1&type=pdf>
- Fowlkes, J.E., Dwyer, D.J., Oser, R.L., Salas, E.: Event-Based Approach to Training (EBAT). *Int. J. Aviat. Psychol.* **8**(3), 209–221 (1998). <https://doi.org/10.1207/s15327108ijap0803>
- Grossman, R., Salas, E.: The transfer of training: what really matters. *Int. J. Train. Dev.* **15**(2), 103–120 (2011). <https://doi.org/10.1111/j.1468-2419.2011.00373.x>
- Harris, D.: *Human Performance on the Flight Deck*. CRC Press, Boca Raton (2012)
- Helmreich, R.L., Klinec, J.R., Wilhelm, J.A.: Models of threat, error, and CRM in flight operations. In: *Proceedings of the Tenth International Symposium on Aviation Psychology*, pp. 677–682 (1999)
- Helmreich, R.L., Merritt, A.C., Wilhelm, J.A.: The evolution of crew resource management training in commercial aviation. *Int. J. Aviat. Psychol.* **9**(1), 19–32 (1999). https://doi.org/10.1207/s15327108ijap0901_2
- Hollnagel, E.: Is safety a subject for science? *Saf. Sci.* **67**, 21–24 (2014). <https://doi.org/10.1016/j.ssci.2013.07.025>
- IATA. *Evidence-Based Training Implementation Guide* (2013)
- IATA. *Data Report for Evidence-Based Training* (2014). <http://www.iata.org/whatwedo/ops-infra/itqi/Documents/data-report-for-evidence-basted-training-aug2014.pdf>
- ICAO. *Procedures for Air Navigation Services - Training* (Doc 9868). Plana (2006)
- ICAO. *Manual of Evidence-based Training* (2013). <http://www.icao.int/SAM/Documents/2014-AQP/EBTICAOManualDoc9995.en.pdf>
- Kanki, B., Helmreich, R., Anca, J.: Crew Resource Management. *Crew Resource Management*, pp. 1–5. (2010). <http://www.scopus.com/inward/record.url?eid=2-s2.0-84882499276&partnerID=40&md5=4aec85e9c8960fc3d3629013edeb580b>
- Kanki, B.G., Greaud, V.A., Irwin, C.M.: Communication variations and aircrew performance. *Int. J. Aviat. Psychol.* **1**(2), 149–162 (1991). <https://doi.org/10.1207/s15327108ijap0102>
- McDonnell, L.K., Jobe, K.K., Dismukes, R.K.: *Facilitating LOS Debriefings: A Training Manual*, March 1997
- Orlady, H.W., Orlady, L.M.: Human factors in multi-crew flight operations. *Aeronaut. J.* **106** (1060), 321–324 (2002)
- Osgood, C.E.: The similarity paradox in human learning: a resolution. *Psychol. Rev.* **56**(3), 132–143 (1949). <https://doi.org/10.1037/h0057488>

- Rankin, A., Woltjer, R., Field, J., Woods, D.: “Staying ahead of the aircraft” and Managing Surprise in Modern Airliners. In: Proceedings of the 5th Resilience Engineering Association Symposium, pp. 209–214 (2013). <http://www.resilience-engineeringassociation.org/download/re-sources/symposium/symposium-2013/>
- Rosen, M.A., Salas, E., Wu, T.S., Silvestri, S., Lazzara, E.H., Lyons, R., Weaver, S.J., King, H. B.: Promoting teamwork: an event-based approach to simulation-based teamwork training for emergency medicine residents. In: Academic Emergency Medicine, vol. 15, pp. 1190–1198 (2008). <https://doi.org/10.1111/j.1553-2712.2008.00180.x>
- Salas, E., Rosen, M.A., Held, J.D., Weissmuller, J.J.: Performance measurement in simulation-based training: a review and best practices. *Simul. Gaming* **40**(3), 328–376 (2008). <https://doi.org/10.1177/1046878108326734>



Now You See It, Now You Don't: A Change Blindness Assessment of Flight Display Complexity and Pilot Performance

Claire McDermott Ealding and Alex Stedmon^(✉)

Centre for Mobility and Transport, Coventry University, Coventry, UK
aviationhumanfactors@icloud.com

Abstract. Synthetic Vision Systems (SVS) provide a revolutionary new technology for modern aircraft flight decks, changing the way pilots see the world by merging a high-resolution representation of their immediate terrain and surroundings underneath the traditional primary flight instruments. Despite its operational benefits, there may be challenges to the effective use of SVS and little research has focused on pilot performance measures. Using custom designed flight display images and a novel Flicker Paradigm, an experiment was designed to measure pilot response time to visual cues on both SVS and conventional electronic displays and also for different levels of pilot experience. Results indicated that change detection was impaired with the SVS display across the pilot ranks. Pilots were typically seven seconds slower and made more errors using the SVS display, supporting other research that suggests that the background complexity of SVS hampers the speed and accuracy of identifying visual cues. Contrary to what was expected, first officers performed both quicker and more accurately than captains. Perhaps this signals the first signs of a new crop of pilots who have been trained using 21st century synthetic and electronic flight displays in today's light training aircraft.

Keywords: Avionics · Synthetic vision systems · Change blindness
Pilot performance · Flicker paradigm

1 Introduction

1.1 New Technologies for New Cockpits

The safe operation of a modern aircraft relies largely on pilots interacting with complex visual displays, through which much of their flight information is presented. These aircraft use primary flight display (PFD) or head-up display (HUD) technologies to present this critical information such as aircraft airspeed, heading, course, and altitude. However, a revolutionary new technology entering the modern flight deck is the development of Synthetic Vision Systems (SVS), which has the capability to merge a computer-generated image of the outside world under that of the pilot's traditional primary flight instruments. At its heart, SVS is driven by an elaborate database of topographical and cultural information that gives pilots a 2D synthetic image of the outside world (i.e. visually the lay of the land). By leveraging the power of Global Positioning

Systems (GPS) and inertial reference systems, SVS displays depict the outside world in real-time, as an aircraft makes its journey from departure to destination (Fig. 1).

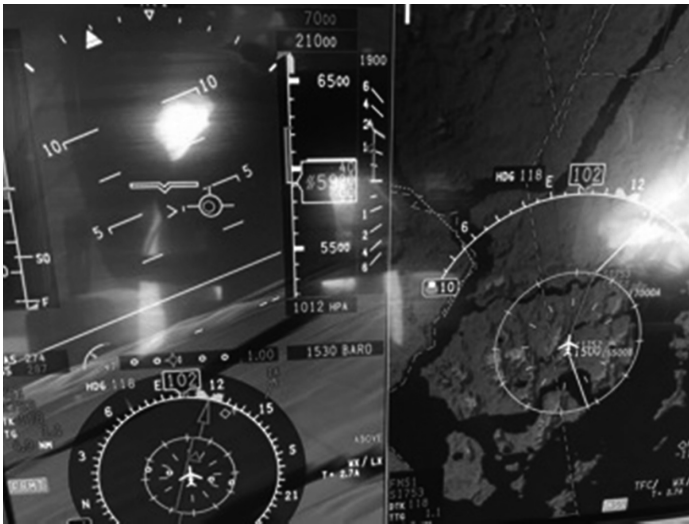


Fig. 1. The Collins Proline Fusion SVS PFD (and SVS moving-map to the right)

Advances in GPS and operational protocols for lower minima approaches have created a positive climate for SVS. Having the added element of SVS data superimposed onto the PFD embodies the cues that are available to pilots in clear weather conditions. Research shows that modern electronic flight displays (particularly those incorporating SVS) have led to more efficient instrument scanning, more informed aeronautical decision making, and improvements in situation awareness (SA) [1–3].

An early demonstration of information visualisation supporting improved SA was performed by taking tabular data of thunderstorms and transforming it into 3D images. Users were able to understand much more about the forthcoming storms by looking at a few seconds of graphical information rather than spending much longer looking at a variety of data in textual formats [4]. By combining disparate data sources into a unified graphical presentation, findings such as these suggest that there is potential for SVS to support pilots in maintaining aircraft state awareness and ‘outside world’ awareness with markedly less effort.

Another application of this technological advance is in relation to how GPS is revolutionising landing procedures into airports surrounded by significant terrain, allowing pilots to fly curved approaches, rather than conventional linear (i.e. straight-in) approaches. ‘Pathway in the sky displays’ are the recommended method for safely flying these complex approaches [5] yet these approaches will eventually become contingent on having SVS. Whilst there is currently no operational advantage to be gained from SVS - such as allowing crew to fly to lower minima - it is a response to many safety issues characterised by real-world events. Thus, another prospect is that

by using information visualisation, SVS might help reduce one of the most pervasive causes of aviation accidents to date: Controlled Flight into Terrain (CFIT) [6, 7].

1.2 Implications for Cognitive Performance

Flight displays design should provide pilots with the tools to support state change detection and maintain SA. Bringing another tool into the flight deck needs to be done in sympathy with already well-established operational protocols and work patterns otherwise it might compromise our natural cognitive capabilities. It is important to understand new technologies from a systems perspective, “new functionality and new technology cannot simply be layered onto previous design concepts because the current system complexities are already too high. Better human-machine interfaces require a fundamentally new approach.” [7, p. 7].

Research using a combination of human and performance modelling shows SVS can be introduced with minimal impact to current pilot scan patterns. Trials revealed that pilot dwell time on the PFD and Navigation Display (ND) deteriorated when SVS was presented on a separate screen. Whereas, performance-modelling trials alone determined that combining the PFD with SVS provided the best configuration, requiring fewer glances away from the primary flight instruments [8]. Nevertheless, one must approach these results with caution, since findings were based on a simplified representation of human performance as demonstrated by a computer, which may not wholly represent the real-world dynamics of human behaviour.

Empirical data from witness testimony, computer gaming accuracy and automobile safety, show that even changes within the direct field of vision can be misinterpreted or even missed entirely (a phenomenon known as ‘change blindness’) when presented with too much surrounding detail [9, 10]. Research has also shown that people have great difficulty simultaneously attending to two visually superimposed scenes [11–13]. Similarly, simply increasing background colour, hues and depth has been found to impair visual processing [14]. Visual perception has been proven to be quite vulnerable to these effects, yet in the aviation domain, little research has been conducted into the operational practicality of SVS and aspects of change-blindness.

Change blindness refers to the difficulty people have in readily detecting changes outside the very small region of focussed attention [15, 16]. A fundamental principle of change detection is that it is a daily problem that people simply cannot attend to all the objects around them [17]. An object must be attended to, to be seen to change [17, 18]. The ability to see in high definition is restricted to a very small area around the focal point of the eye, and so perceiving detail in the environment is incumbent on effective and frequent eye movements [19]. In a simple scene, this does not take much effort, but increasing the amount of information vying for one’s attention will place increased cognitive demands on the observer, thus slowing the visual search or increasing opportunities for them to miss information or state changes in the visual scene.

Previous research has demonstrated that there is greater likelihood of a failure to integrate information across saccades, distractions or ‘mud splashes’ or ‘Flicker’ episodes (i.e. techniques used to investigate change blindness in laboratory conditions) compared with changes introduced during direct fixation [19, 20]. Furthermore, how meaningful information might be is a key factor in determining what people decide to

process, and thus becomes an important factor in change blindness [19]. As a whole, what these findings illustrate is that just because one's eyes are open, does not necessarily mean everything is being seen.

A conclusion from the National Weather Service (NWS) best illustrates the argument for SVS - the use of numbers requires analysis, but the use of imagery induces intuition [4]. However, although enough evidence exists for the merging of SVS with the PFD, decades of research challenges the practicality of this system. Naturalistic scenes interfere most with target searches [10, 21] and changes are masked by the sheer amount of colour or detail on a display [22].

Essentially, having a highly accurate depiction of the outside world could be considered intuitive, but increasing the colour and dynamics within these displays could also have an adverse impact on a pilot's ability to notice changes or transitions, and sustain attention to other key visual areas.

1.3 Rationale

A virtual SVS depiction of the external environment may be intuitive and improve situation awareness by helping pilots to scan their displays more efficiently. However, despite this operational benefit, relatively little research has explored the relationship between sophisticated SVS displays and pilot performance.

Building on previous research into change blindness, the aim of this study was to investigate any response time difference for pilots detecting visual changes occurring on a conventional flight display compared with an SVS superimposed PFD. A further element of the research explored the topic of change detection performance as a function of expertise. It was predicted that changes occurring within conventional PFDs would be detected quicker than changes occurring within SVS PFDs (since increasing detail, motion and colour of backgrounds have been shown to hamper visual detection speed of cues) and due to more developed scanning techniques and increased knowledge associated with greater flight time, captains would be quicker at detecting changes across both conditions than first officers.

2 Method

Using custom designed stimuli that simulated both conventional (non-SVS) and SVS flight displays, an experiment conducted to measure pilot response time to visual cues (as a surrogate measure for SA).

2.1 Participants

An opportunistic sample of 18 pilots consisting of nine captains (age 34 to 52 years) and nine first officers (age 23 to 34 years) who were qualified in both traditional as well as SVS avionics. Five captains and one first officer had over 5,000 h flying time: four captains and five first officers had 1,500 h to 5,000 h flying time; three first officers had less than 1,500 h flying time.

2.2 Apparatus

The experiment was presented on a Macintosh Apple MacBook Air 13-inch laptop. Some participants conducted the experiment remotely on their own laptop computer. A 'Flicker Paradigm' was developed using Inquisit Software with the custom-made images. Participants generally positioned themselves as close as possible to the seating position they took in the cockpit. General flight deck anthropometrics have pilots seated anywhere between 500 mm to 700 mm from the display, with the screen in the centre of the visual field (within 30° of binocular view) and approximately 15° below the normal line of sight [23].

2.3 Design

The study employed a 2×2 mixed design. The between groups factor was 'experience' (captain vs first officer) and the within-groups factor was 'display' (conventional vs SVS). Dependent variable measures were collected for response time data to on-screen changes in the flight instrumentation displays.

A counter-balanced repeated measures approach ensured that the order of the stimuli was randomised for each participant to remove any learning, practise or fatigue effects.

The Flicker Paradigm is an adaption of tests that simply present an original and modified image back to back. By inserting a blank screen between each image (i.e. a 'flicker'), fixated attention would be required to notice any changes and coincident with a natural blink since pilots do not typically dwell on the PFD for more than 30–40 s at a time [24], or a simultaneous on-screen change to an icon [25].

Images simulating flight displays were designed in accordance with the AC25-11A/B and DEF-STAN standard for PFDs, and presented to the participants in pairs (original and modified). Each image pair was identical, except for one single change (Figs. 2 and 3). A 'change' was defined by; an object disappearing or reappearing, an object changing colour, an object changing position, or an alphanumeric value change. These changes were only incorporated into the flight symbology that a pilot would normally be expecting to change (i.e. a colour of a speed read-out was not changed to a colour that would be inconsistent with reality). Although the changes are highlighted for ease of identification in Figs. 2 and 3, they were not in the experiment and so participants had to judge for themselves what change might have occurred.

The original image was repeatedly alternated with a modified image (240 ms each), separated with a Inter-Stimulus Interval (ISI) (a blank screen) lasting 80 ms. The images and ISI's were alternated until the participant detected the change or 60 s had passed [26], whichever came first. The sequence was looped as 'original' > 'original' > 'ISI' > 'modified' > 'modified' > 'ISI'. This created a degree of uncertainty about when stimulus change might occur [18] and also gave participants more of an opportunity to process each image. This made the search more naturalistic and participants were less able to predict an oncoming change.

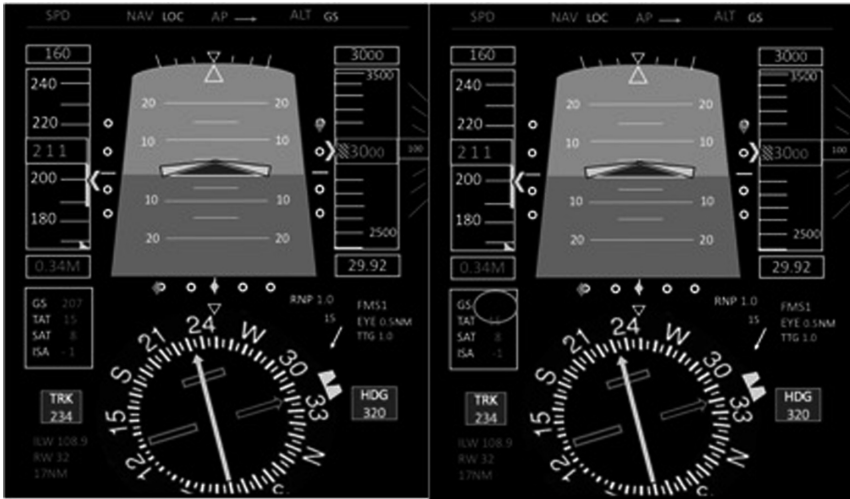


Fig. 2. Ground Speed reading change between left and right displays (highlighted on right hand display in bottom left panel for ‘GS’)

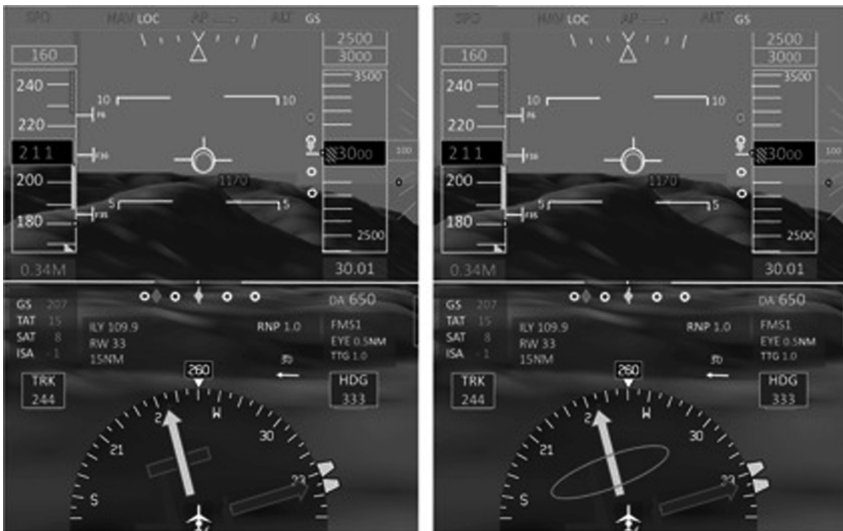


Fig. 3. Localiser displacement indication change between left and right displays (highlighted on right hand display under centre white arrow)

2.4 Procedure

Participants were shown a total of twenty pairs of images presented across the two conditions: 10 pairs of conventional PFDs without SVS and 10 pairs of PFDs with simulated SVS information. Each pair was identical, except for one change.

Participants were tasked with detecting the change. Pushing the spacebar when the change was detected automatically logged their response time while also triggering the next image pair in the sequence. Immediately after each image pair the participant was asked to report the type of change they saw by typing their observations into a dedicated text box. This was done to help ensure accuracy and prevent against random clicking.

3 Results

Data from 18 participants were included in the final analyses. Within this sample 50% were captains ($N = 9$) and 50% were first officers ($N = 9$). Analysis of Variance tests were conducted on the response time data.

3.1 Response Time as a Function of Display Type

A significant main effect for 'display' was observed ($F = 4.9/F_{crit} = 4.15$, $p = 0.034$) illustrating that display complexity influenced response time. Pilots were slower at detecting changes occurring on the SVS display ($M = 25$ s, $SD = 9$ s) than Conventional displays ($M = 19.5$ s, $SD = 7$ s). From the sample, 83% of pilots were faster at detecting changes on the Conventional display, whereas just three participants were quicker in the SVS condition. For changes detected in the first 40 s pilots noticed changes in 89% of the trials for the conventional display compared with only 78% for the SVS condition.

3.2 Response Time as a Function of Pilot Expertise

A significant main effect for 'experience' was observed ($F = 6.13/F_{crit} = 4.15$, $p = 0.019$), illustrating that experience affected the response times for changes on the displays. However, this was in a different direction than anticipated. Less experienced first officers were on average 6 s faster, and more accurate (i.e. they noticed more changes in the displays) than the more experienced captains. Captains missed 17% of the changes presented to them while first officers missed only 8% of the same changes. For changes detected within the first 40 s, captains noticed changes in 78% of the trials and first officers noticed changes in 89% of the trials. Overall co-pilots demonstrated quicker change detection than captains on both display types. This was reflected in results for an interaction (display x experience) although this was not statistically significant ($p = 0.8$).

4 Discussion

4.1 Response Time as a Function of Pilot Expertise

The goal of SVS is to help pilots make more informed decisions by providing them with a clearer picture of their surroundings. The purpose of this study was to explore if these intricate displays increased the pilots' information processing burden.

Specifically, if the increased detail in the displays hampered their ability to scan the flight instruments effectively and notice critical changes in the information before them.

Findings from prior research has collectively found that people tend to become distracted and overburdened by increased background detail [22]. This study hypothesised that the complexity within the flight display would negatively impact on change detection speed. This appeared to be supported by the results as significantly slower detection times were observed to changes occurring on the SVS displays.

Although the results from this experiment do not refute that imagery makes for a more efficient and intuitive scan, they do indicate that small on-screen changes can be masked. Yet, while the primary goal of this research was to explore some of the pitfalls of the new SVS display designs, it also brought to light some of the underlying human factors issues unique to corporate aircraft operators. The misguided universal assumption that pilot expertise can be indexed by age and hours, and how multiple users from an array of ages, generations and ranks are being expected to operate highly sophisticated, un-customisable technology.

An important cognitive skill for being an effective crewmember is the ability to scrutinise a myriad of information in a timely fashion. Yet an important finding was that captains were generally slower than first officers at spotting changes in both display formats (i.e. conventional and SVS). It may have been possible to surmise that careful, knowledge-driven searches are characteristic of the more senior ranking and experienced crewmembers, which might explain the longer detection times. However, one clue to understanding why the first officers excelled in the change detection tasks could be explained by the renovation of the fleet of training aircraft around the world. Over the last decade, virtually all new aircraft are manufactured with some degree of digital (glass) cockpit technology (Fig. 4, left-hand image). Many of the newer pilots may never have even flown with traditional analogue gauges as the more experienced captains may have done in the past (Fig. 4, right-hand image).



Fig. 4. Glass cockpit display (left) and traditional analogue display (right)

4.2 New Pilots for New Cockpits

The new generation of training aircraft might, in part, be the reason for the generation of new first officers performing better on digital displays (i.e. conventional *and* SVS displays) than their captains. In this situation, expertise might be more attributable to circumstance and contextual learning (i.e. being exposed to technology throughout one's early life and training) and less a function of hours and rank.

Taking that same logic, captains, with the foundation of their skills and proficiency honed using analogue and conventional displays, could find themselves coping with distractions from the additional information and learning unfamiliar displays. This view is consistent with research into skill acquisition, that when knowledge is first acquired it is organised in a way that, thereafter, can be accessed automatically through pattern-based retrieval [27, 28]. The cues available to captains on newer electronic displays may not complement their pre-existing patterns or mental models, thus requiring a period of readjustment or acclimatisation to better interact with SVS displays.

These findings resonate with the statement, "too often, when a new system is introduced, it is assumed that trainees are already experts in the processes the system is intended to monitor and control" [29, p. 444]. A pilot who is proficient and can scan their instruments in a largely automatic fashion, will not naturally be equally proficient with a highly advanced SVS display. Thus, it is conceivable that captains performed worse because they had to commit more attentional resources and expend more effort engaging in active learning, in order to overcome their pre-existing mental models. Perhaps this reveals a generational challenge that new computerized technology might not be able to be embraced by all users as a 'one-size-fits-all' solution. As with other sophisticated devices, such as mobile phones and computers [30] changing technologies on flight decks will need to be tailored to the needs, requirements and limitations of a wider selection of users.

While the findings of this research provide a tantalising insight into the way that new technologies may emerge in new applications, it is important to bear in mind that the sample population in this study was not representative of the pilot population at large. As such, the variance in performance between captains and first officers may need to be explored and verified with further research. To better simulate the tasks and workload that accompany the normal flight regimen, future improvements to this experiment can be achieved by carrying out SVS change blindness trials in simulator environments while also using a more industrywide representative crew sample.

Within commercial transport operations, for instance, generally the captain and the first officer will take control of alternate flights, getting roughly an equal number of take-offs and landings. Whereas, in charter aviation, the first officer may be restricted from acting as the 'pilot flying' until they have logged several hundred hours of flight time. Depending on proficiency, in some rare cases, first officers could spend the better part of a year operating only as 'pilot monitoring' logging only a handful of take-offs and landings. This implies that they may have developed an affinity for scanning displays, which may have contributed to their increased change detection performance. Finally, there may also have been a lack of engagement among the more senior crew

members (i.e. captains) and perhaps an increased engagement among the junior crew members (i.e. first officers) who were more conscious of their performance.

5 Conclusion

Overall, these findings leave us with the view that complex and sophisticated systems can become a challenge, even for experienced users. By providing a graphical depiction of the outside world environment, SVS can augment pilot SA and aircraft state awareness beyond anything previously introduced; and be necessary for more advanced approaches into precipitous terrain. However, all too often with rapidly developing technologies, the potentially negative side-effects are excluded from initial exploratory studies. Two strongly counterintuitive results arise from the findings of this experiment, indicating that there is perhaps another side to the coin with SVS.

SVS presents a tantalising and unsurpassable method of providing situation awareness. Therefore, it cannot be reasonably placed on a separate display without causing significant decrements in visual scanning. However, given the potential risks associated with decrements in change detection, there may be a requirement for SVS to be assuaged during phases when distraction would be most critical.

Also brought to light was how expertise is not only characterised by hours, but by the ability to adapt to constantly evolving technologies. The first officers performed their tasks more quickly and more accurately than captains. This was unexpected considering that experts (based on hours and rank) were purported to conduct more effective scans. Maybe this denotes a failure to consider the way SVS equipment might interact with the already established behaviours and mental models built up through years of exposure on older technologies. Or perhaps pilots need to be more mindful of the way they interact when presented with these powerful new technologies.

Funding. This research was part of an MSc Thesis and therefore not formally funded. The authors acknowledge the support of Zetta Jet Flight Department, specifically Eric Rastler Chief Pilot, for granting approval to conduct the research using company personnel.

References

1. Endsley, M.: Towards a theory of situation awareness in dynamic systems. *Hum. Factors* **37** (1), 32–64 (1995)
2. Endsley, M., Jones, D.: *Designing for Situation Awareness: An Approach to User Centred Design*, 2nd edn. CRC Press, Boca Raton (2016)
3. Foyle, D.C., Kaiser, M.K., Johnson, W.W.: Visual cues in low-level flight: implications for pilotage, training, simulation, and enhanced/synthetic vision systems. In: *American Helicopter Society 48th Annual Forum*, vol. 1, pp. 253–260 (1992)
4. Wilhelmson, R.B., Jewett, B.F., Shaw, C., Wicker, L.J., Arrott, M., Bushell, C.B., Bajuk, M., Thingvold, J., Yost, J.B.: A study of the evolution of a numerically modelled severe storm. *Int. J. Supercomput. Appl.* **4**(2), 20–36 (1990)

5. Medjal, S., McCauley, E., Beringer, D.: Human Factors Design Guidelines for Multifunction Displays. U.S. Department of Transportation Office of Aerospace Medicine: Washington (2001)
6. Priznell III, L., Kramer, L., Bailey, R., Arthur, J., Williams, S., McNabb, J.: Augmentation of Cognition and Perception Through Advanced Synthetic Vision Technology. NASA Langley, Hampton (2005)
7. Theunissen, E.: Integrated Design of a Man-Machine Interface for 4-D Navigation. Delft University Press, The Netherlands (1997)
8. Deutsch, S., Pew, R.: Examining new flight deck technology using human performance modeling. In: Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, pp. 108–112. Human Factors and Ergonomics Society, Santa Monica (2004)
9. Wolfe, J.: What can 1 million trials tell us about visual search? *Psychol. Sci.* **9**(1), 33–39 (1998)
10. Caroux, L., LeBigot, L., Vibert, N.: Impact of the motion and visual complexity of the background on players' performance in video game-like displays. *Ergonomics* **56**(12), 1863–1876 (2013)
11. Neisser, U., Becklen, R.: Selective looking: attending to visually specified events. *Cogn. Psychol.* **7**, 480–494 (1975)
12. Fischer, E., Haines, R.F., Price, T.A.: Cognitive Issues in Head-Up Displays, NASA Technical Paper 1711. NASA Ames Research Centre, Moffett Field (1980)
13. Stedmon, A.W., Kalawsky, R.S., Hill, K., Cook, C.A.: Old theories, new technologies: cumulative clutter effects using augmented reality. In: IEEE International Conference on Information Visualisation 1999: International Conference on Computer Visualisation, 14–16 July, London, U.K. (1999)
14. Wolfe, J.: Visual attention. In: De Valois, K.K. (ed.) *Seeing*, 2nd edn, pp. 335–386. Academic Press, San Diego (2000)
15. Grissinger, M.: Inattention blindness: what captures your attention? *Pharm. Ther.* **37**(10), 542–555 (2012)
16. Simons, D.: Current approaches to change blindness. *Vis. Cogn.* **7**(1–3), 1–15 (2000)
17. Rensink, R.: The dynamic representation of scenes. *Vis. Cogn.* **7**, 17–42 (2000)
18. Rensink, R., O'Regan, J.K., Clark, J.: To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci.* **8**(5), 368–373 (1997)
19. Simons, D., Levin, D.: Failure to detect changes to people during a real-world interaction. *Psychon. Bull. Rev.* **5**(4), 644–649 (1998)
20. Henderson, J., Hollingworth, A.: Global transsaccadic change blindness during scene perception. *Psychol. Sci.* **14**(5), 493–497 (2003)
21. Caroux, L., LeBigot, L., Vibert, N.: Impairment of shooting performance by background complexity and motion. *Exp. Psychol.* **62**(2), 98–109 (2015)
22. Wolfe, J., Oliva, A., Horowitz, T.S., Butcher, S.J., Bompas, A.: Segmentation of objects from backgrounds in visual search tasks. *Vis. Res.* **42**, 2985–3004 (2002)
23. DEFSTAN 00-25: Human Factors for Designers of Equipment. Crown Copyright: Ministry of Defence Directorate of Standardization, Glasgow (1992)
24. Mumaw, R., Sarter, N., Wickens, C.: Analysis of pilots' monitoring and performance on an automated flight deck. In: 11th International Symposium on Aviation Psychology, Ohio State (2001)
25. Rensink, R.: When good observers go bad: change blindness, inattention blindness and visual experience. *Psyche: Interdisc. J. Res. Conscious.* **6**(9) (2000). <http://cogprints.org/1050/3/psyche-6-09-rensink.pdf>. GoogleScholar. Accessed 09 Feb 2018
26. Rensink, R.: Visual search for change: a probe into the nature of attentional processing. *Vis. Cogn.* **7**, 345–376 (2000)

27. Fitts, P.M., Posner, M.I.: *Human Performance*. Brooks/Cole Publishing Co., Belmont (1967)
28. Ericsson, K., Lehmann, A.: Expert and exceptional performance: evidence of maximal adaption to task constraints. *Ann. Psychol. Rev.* **47**, 272–305 (1996)
29. Durlach, P.: Change blindness and its implications for complex monitoring and control systems design and operator training. *Hum.-Comput. Inter.* **19**, 423–451 (2004)
30. Pattison, M., Stedmon, A.W.: Inclusive design and human factors: designing mobile phones for older users. *PsychNology J.* **4**(3), 267–284 (2006)



Experimental Evaluation of a Scalable Mixed-Initiative Planning Associate for Future Military Helicopter Missions

Fabian Schmitt^(✉) and Axel Schulte

Institute of Flight Systems, University Bundeswehr Munich,
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
{fabian.schmitt,axel.schulte}@unibw.de

Abstract. This article describes a scalable mixed-initiative planning concept, in which a human pilot is assisted during mission (re-)planning by an artificial planning agent. The agent serves as an additional team member and enables rapid planning and re-planning of multiple vehicles. For this purpose, the agent adapts its extent of assistance based on the necessity of the given situation. The concept was implemented for the use case of manned-unmanned teaming in future military helicopter missions. Thereby, the mixed-initiative agent was implemented with three different levels of automation. The article focuses on the experimental evaluation with German military helicopter pilots. Results show the advantages of the scalable mixed-initiative concept especially in time critical and high workload situations.

Keywords: AI systems · Associate systems · Mixed-initiative Problem solving

1 Introduction

The mission planning, re-planning and management of multiple vehicles in dynamic scenarios by a single pilot is a highly topical field of research. A major issue within this research field is the risk for excessive operator *mental workload (MWL)* and loss of operator *situation awareness (SA)*. Systems for partly or fully automated planning and scheduling can solve complex multi-vehicle problems in reasonable time. However, these systems are often not directly suitable for use in user-oriented incremental and collaborative planning [1]. Instead, such systems are developed to perform their dedicated function. This might result in a number of fundamental human factors related problems. A loss of *situation awareness (SA)* and *plan awareness (PA)* is likely to occur in critical situations, because the operator is not integrated into the planning process anymore. Furthermore, complacency and over-reliance may occur, because operators may overly trust those planning systems.

In order to counteract such issues, we propose a *mixed-initiative (MI)* planning approach. We define *mixed-initiative* as a cooperative approach between a human operator and a cognitive agent to solve a common planning problem. Thereby, both the

human operator and the agent can take initiative over the planning process and direct the process into a certain direction.

In this article, we present a *scalable mixed-initiative* planning agent. This agent is able to adapt the extent of its intervention to the situation. The emphasis of this article is on its experimental evaluation. In the following two sections, we will describe firstly our application and secondly the concept of this agent. The subsequent section describes an experimental campaign, which was conducted recently with German military helicopter pilots to evaluate the agent in complex mission scenarios. The article concludes with results of that research campaign.

This work builds on research originated in [1]. Multiple other *mixed-initiative* approaches were already developed in [2–4] for various domains. However consequent experimental evaluations are rare.

2 Application

In our research, we look at *Manned-Unmanned Teaming (MUM-T)* in the field of military helicopter missions. Here, a manned two-seated helicopter is teamed with three small Unmanned Aerial Vehicles (UAVs). The UAVs serve as detached sensor platforms. They are designed to conduct route- and area reconnaissance and are able to detect safe landing points in hostile environments. Operational benefits include an increased sensor range of the manned platform, a better understanding of the tactical situation, increased lethality for the human pilots and eyes on target capabilities. The unmanned systems are controlled directly out of the cockpit by one of the pilots to reach a high level of interoperability. Thereby, pilots can access new reconnaissance information instantaneously and react rapidly if required. In our concept, the pilot in command (PIC, non-flying) serves as battlefield manager. He is responsible for the tactical planning and re-planning of the manned/unmanned team during the mission. The pilot's planning task consists of helicopter route planning (primary route and alternative route), identification and assignment of reconnaissance tasks to the UAVs and temporal coordination between different vehicles. Therefore, he works cooperatively with our *mixed-initiative* planning agent. The agent proposes future helicopter routes and UAV tasks, identifies threat conflicts in the current plan and offers suitable solutions. Furthermore, it helps optimizing a given plan. The communication between pilot and agent is dialog based. When the agent comes up with a recommendation, the pilot can either accept, reject or ignore the recommendation.

3 Concept

3.1 Work System

We will use a formal graphical description method, designed for *Human-Autonomy Teaming (HAT)* [5], to present the *mixed-initiative* approach. This method was developed our institute in order to structure top-level *HAT* designs and is based on the work process. Based on a given work objective and the corresponding work

environment, a certain work output shall be generated. The work system differentiates between the worker on the left-hand side and the tools on the right-hand side. The worker has knowledge of the overall work objective and tries to reach that objective by own initiative. The human worker is furthermore authorized to change this work objective by own initiative. To achieve that work objective, the worker uses given tools which are shown on the right-hand side of the work system. The tools are subordinates, i.e. hierarchically degraded with respect to the worker. More general information about the work process and work system notation can be found in [5]. The conceptual design of our *mixed-initiative* work system is presented in Fig. 1. On the worker (left-hand) side is the pilot in command, who is responsible for airborne mission (re-)planning. Three dislocated UAVs (right-hand side) serve as tools and can be used to fulfill the mission objective. A cognitive agent on-board each UAV provides a delegation interface to the pilot to enable task-based guidance and controls the actual UAVs in supervisory control. The human pilot uses degraded planning tools to (re-)plan or modify the mission plan whenever required.

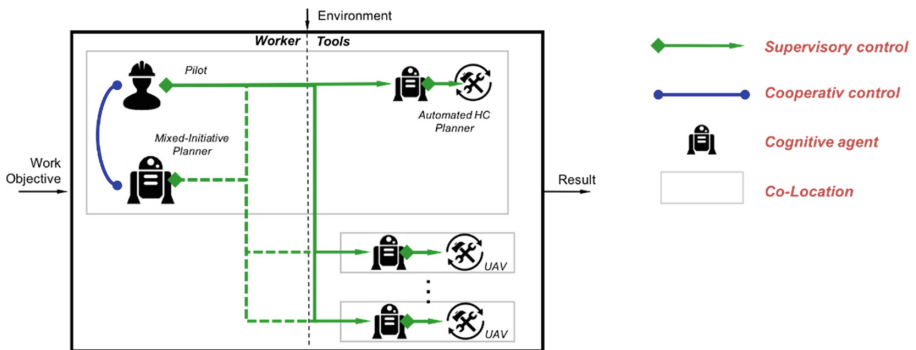


Fig. 1. Work system notation for the *mixed-initiative* approach

Additionally, we added a cognitive agent next to the pilot on the worker side, which represents the *mixed-initiative* agent. As a consequence of the placement on the worker side, the agent has to know the mission objective and can intervene in the planning process by own initiative. Thereby, initially the pilot provides information about a mission goal to the agent. Based on that, the agent has access to the described tools (i.e., UAVs and helicopter route planner) in the same way as the human pilot.

3.2 Agent Behavior

The role of the planning agent is to assist the pilot in the mission planning process. It shall ensure, that a satisficing and valid mission plan is available at all times throughout a mission. As long as the pilot is doing a good job at moderate workload, the agent shall remain passive as there is no need for any assistance. However, when the pilot needs any support, it shall identify future mission tasks as well as conflicts in the current plan and proposes suitable solutions to the pilot. It is important to point out, that

the pilot cannot assign tasks to the planning agent. Rather, the planning agent decides on its own, when to intervene.

We formulated the following behavior rules for the planning agent [6]:

- intervene as little as possible,
- intervene as late as possible (to enable the pilot to solve the problem on his own, if possible),
- leave as much work to the pilot as possible to maintain pilot's SA (i.e. adapt the extent of assistance/intervention to the pilot's workload),
- intervene incremental, rather than complex, whenever possible (a complex proposal might be hard to understand, *plan awareness* might diminish).

When designing the agent's behavior, we also oriented on Herbert Simon's *Satisficing Principle* [7] from organization psychology which describes the behavior of human decision makers. Simon emphasized especially two points:

1. The human in general does not try to find an optimal solution to a given problem. Instead, he stops working on the problem as soon as the solution is sufficient for him.
2. Sufficient quality depends on the decision maker's personal level of aspiration. The level of aspiration is thereby determined by the criticality of the given situation.

Based on Simon's principle our agent shall weigh if an interaction is really required in a given situation. Also, in low workload situations, the pilot might have an increased level of arousal. We formulated the following additional behavior rules:

- aspire after a sufficient rather than an optimal plan (the pilot does not aspire after an optimal plan), and
- adapt the aspiration level to the current tactical situation, and the pilot's mental workload.

3.3 Scalable Mixed-Initiative

To be able to adapt the extent of intervention to the pilot's *mental workload*, we defined three *Levels of Automation (LOAs)*. Depending on the given situation, one of the levels might be most suitable. The three levels of automation are:

1. *Silent mode*: In this mode the agent is designed to leave all work to the pilot. It corrects only errors in the plan (low workload mode, only available on initial mission planning).
2. *Incremental planning mode*: 1. + the agent takes initiative on the expansion of the plan using single/incremental steps of interventions. (medium workload mode, basic mode during flight) The agent shall reduce the share of work of the pilot during the planning process.
3. *Extensive planning mode*: 2. + the agent intervenes with complex/extensive interactions comprising multiple planning steps at a single time. (high workload mode, only in critical situations with high task load) In this mode, the agent is able to propose task sequences for multiple aircrafts at the same time to achieve rapid plan progress.

We defined a default level respectively for initial planning (*silent mode*), re-planning due to changes of tactical situation and throughout the flight (*incremental planning mode*), and re-planning due to changes of the mission objective (*extensive planning mode*). If there are no further information available about the pilot's *mental workload*, one of these levels shall be chosen automatically depending on the given situation. However, ideally, in *scalable mixed-initiative* the extent of agent interventions (i.e. the level of automation) shall be adapted automatically to the pilot's current workload.

While *MWL* can, in general, be reduced by increasing the extent of an intervention, pilot's *situation-* and *plan awareness (PA)* might probably diminish. The probability for lacks in *plan awareness* increases, because the pilot's part in the cooperative planning process decreases. Balancing *PA* and *MWL* is therefore one of the crucial factors in *scalable mixed-initiative*.

3.4 Distribution of Tasks Between Pilot and Agent

In order to balance workload, task responsibilities can be shared between pilot and agent by default. In our application, the pilot shall be responsible for the planning primary helicopter routes and the task identification and assignment to UAVs. In the contrary, the planning agent shall be responsible for planning of alternative helicopter routes as well as scheduling (permanently substituting assistance). However, if the pilot is not satisfied with the agent's solution, he can modify it.

4 Functional Architecture and Implementation

This chapter briefly describes the implementation of the *mixed-initiative* agent in the helicopter mission simulator at the Institute of Flight Systems.

4.1 Mixed-Initiative Agent

Four key capabilities are required to achieve the desired supporting agent behavior. These are described in the following:

Capability for Reasoning and Planning in the Given Application Domain. This capability is used to correct flaws in the current plan or to determine future actions, which are required in order to reach the mission objective. Based on this capability, the agent can compare planning solutions with each other and thus help optimizing the mission plan.

We modelled our planning domain using a-priori knowledge. We used PDDL as modelling language, which is an action-centered language, which is widely used to solve planning problems [8]. Core of PDDL are actions with pre- and post-conditions that describe the applicability and the effects of actions. In our application, these actions

comprise helicopter and UAV specific actions. We use a PDDL planner which works based on our mission domain and a problem file. Additionally, we use a CPLEX planner for rapid task assignment, optimization and scheduling [9].

Capabilities for Activity Determination. Knowledge about the pilot's planning activities is helpful to determine if the pilot is already working on the mission plan or not. If he is already working on an existing flaw in the plan, the agent might not need to intervene anymore. Honecker et al. developed an evidential reasoning approach derived from Dempster-Shafer theory, which is used to infer the current pilot's actions [10].

Capabilities to Estimate Pilot's Workload. Knowledge about the pilot's current workload is used to override the default *Level of Automation* for a given situation in order to adapt better to the pilot's current mental state. Brand and Schulte [11] presented a method and implementation of a workload-adaptive associate system.

Capabilities for Intervention. Finally, the agent needs the ability to interact with the pilot based on the usage of all aforementioned capabilities. Therefore, it uses rules to prioritize planning activities (e.g. the detection of a landing point is more important than the reconnaissance of an alternative helicopter route). Furthermore, the pilot's activities and his workload must be taken into account. Based on that the *mixed-initiative* agent has to infer a course of actions to generate a mission plan in cooperation with the pilot. We used PDDL to model possible interactions and prioritizations as actions with pre- and post-conditions. Costs were assigned to each action to model the *Satisficing* principle.

More details of the implementation can be found in [6, 12].

4.2 User Interface

We developed an HMI that satisfies requirements for mission planning and communication between pilot and agent. The HMI has the following three components:

- a mission interface,
- a tactical map display, and
- a dialog interface.

The mission interface is used by the pilot to specify the mission objective. The tactical map display is used by the pilot to sketch the mission plan and command tasks to the UAVs. Therefore, the pilot uses an object-oriented context menu. The map display is also used by the agent to visualize plan proposals and alternatives (Fig. 2). The dialog interface is used primarily by the agent to communicate with the pilot. The agent uses pushes text messages to propose new tasks or solutions (Fig. 3). Each message contains the particular problem identified by the agent. Additionally, most messages contain a proposed solution for that problem and an option for the pilot to either accept or reject the proposed solution. The solution is also visualized graphically on the map display in magenta.

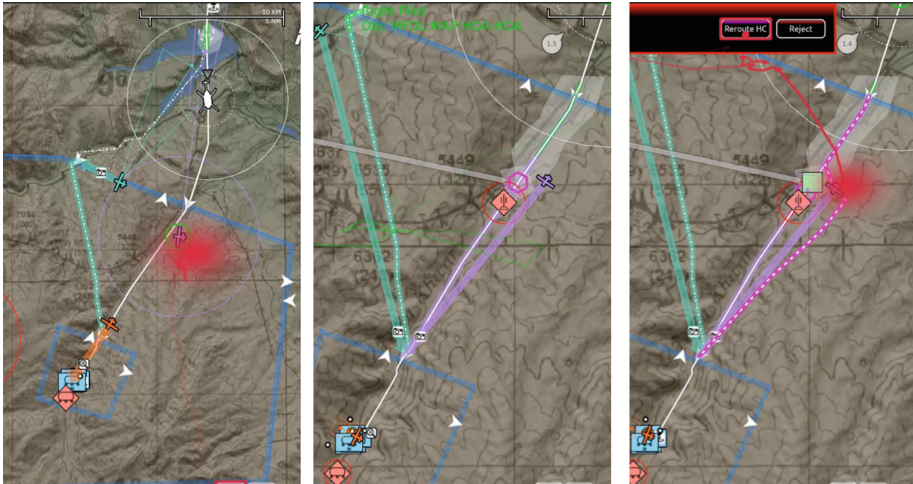


Fig. 2. Initial planning with preplanned alternative (left), pop-up threat on primary route (center) and re-planning due to threat (right). The red spots represent the pilot's gaze position. (Color figure online)

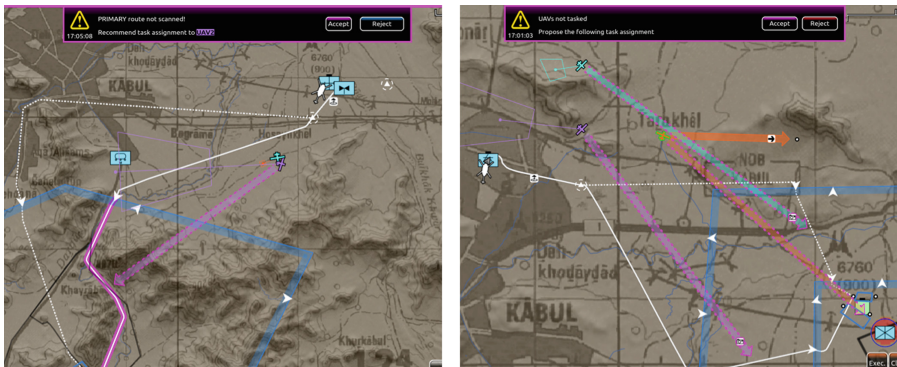


Fig. 3. Incremental task proposal (left) and extensive task proposal (right) of the agent.

5 Experimental Evaluation

5.1 Experimental Setup

A human-in-the-loop experiment was developed in order to evaluate our MUM-T configuration with crews of military pilots. Within this context, the cooperation between the pilot in command and the planning agent for the previously described three levels of automation was also evaluated. The experiment was conducted in the helicopter mission simulator at the Institute of Flight Systems (Fig. 4). In our experimental setup, the pilot non-flying was responsible for mission planning and re-planning of the manned helicopter and additionally 3 three unmanned systems. We exposed all crews

to multiple changes of the tactical situation as well as changes of the mission objective inflight. Thereby, the pilots were assisted by our *mixed-initiative* agent. During the missions, we observed pilot behavior using gaze tracking and manual system interactions). After each mission we gathered subjective ratings using prepared questionnaires.



Fig. 4. Helicopter mission simulator at the IFS

5.2 Participants

Eight military helicopter pilots from the German armed forces participated in the study. The subjects included 1 current and 7 former military pilots from 31 to 59 years of age ($M = 50.4$, $SD = 9.2$). Flight hours varied between 535 and 6850 ($M = 3933$, $SD = 1807$). The pilots were subdivided into four crews of two pilots each. Thereby, one of the pilots served as pilot in command (planning & mission management) while the other one served as pilot flying (responsible for control of the actual manned helicopter). PIC flight hours ranged between 2930 and 4920 h ($M = 3937$, $SD = 804$).

5.3 Scenario

We conducted five military MUM-T helicopter transport missions within three days. Each mission consisted of an initial planning phase on ground, an ingress phase into hostile territory, detection of pop-up threats, at least one landing in hostile territory, a change of the mission objective inflight, and an egress phase. Therefore, each mission contained at least a single use case for each level of automation:

- use case *silent mode*: initial mission planning on ground,
- use case *incremental planning mode*: change of tactical situation (pop-up threat on helicopter route), and
- use case *extensive planning mode*: change of mission objective.

The sequence of missions was kept equal across all crews. The designed scenarios were rated afterwards by each pilot using Likert scales (ranging from 1/negative to 7/positive). Realism of the scenarios was rated very well ($M = 5.6$, $SD = 0.7$). Mission sequence of actions was also rated positive ($M = 5.3$, $SD = 0.83$). Likewise, the simulation was rated rather positive ($M = 5.0$, $SD = 1.0$). Mission duration varied between 31 min and 71 min ($M = 45.77$ min $SD = 10.25$ min). Prior to the first mission, pilots were trained on the simulator for 2 full days. Thereby, the training consisted of a tutorial and 4 training missions with increasing complexity. Learning objectives were tested after completion of the tutorial for each pilot.

6 Results

We analyzed three different situations in each mission, which matched to our use cases (initial planning – silent mode, re-planning due to change of tactical situation, re-planning due to change of mission objective). In the following, we present results of objective measures and subjective pilot ratings. Pilots were asked to rate agent behavior for the described use cases after each mission. The questionnaires were realized using Likert scales ranging from 1/negative to 7/positive, with 4 being neutral. Results will be discussed in the following subsections.

6.1 Use Case: Initial Planning

By default, during initial mission planning, the agent was in the *silent mode*, generating no proposals, since time and workload are no crucial factors in this situation. In case of a detected workload peak, the agent was set into the *incremental planning mode* by the workload adaptive associate system component. In 70% of the initial planning cases, the agent remained in the *silent mode* and did not take any initiative (Fig. 5).

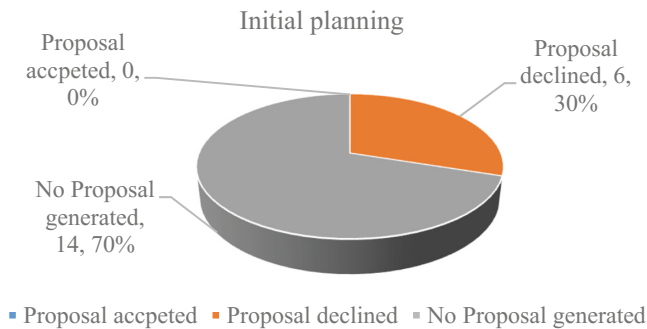


Fig. 5. Results of accepted and declined proposals during initial mission planning phase (N = 20)

In all other cases, the agent’s automation level was changed to *incremental planning mode* due to higher workload. In these cases, the agent initiated at least one incremental proposal. These proposals were either ignored or actively rejected by the subjects.

However, the extent of interventions for initial planning was rated well in general ($M = 0.2$, $SD = 0.75$), (Fig. 7). Subjects stated that they did not like incremental proposals during the initial planning process, once they started the planning process. One pilot stated that he preferred the *extensive planning mode* (ideally a proposal for a full mission plan) at the beginning of this planning phase. The other pilots preferred the *silent mode*. These results indicate, that the silent mode is the best suitable default mode for initial planning. In case of high workload, the agent could directly switch into the *extensive planning mode*. However, more research is required here.

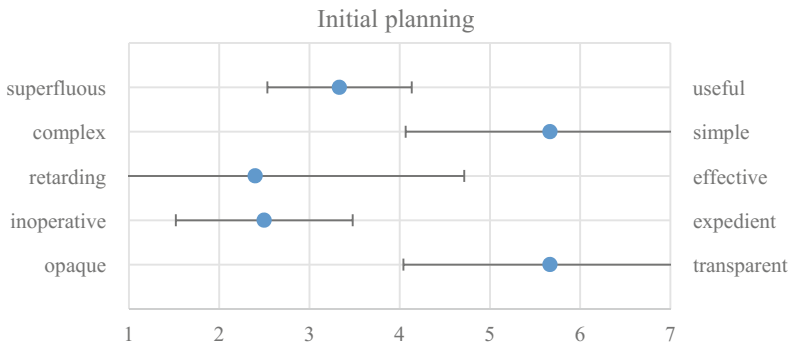


Fig. 6. Results of subjective ratings for interventions during initial mission planning (mean and standard deviation, N = 6)

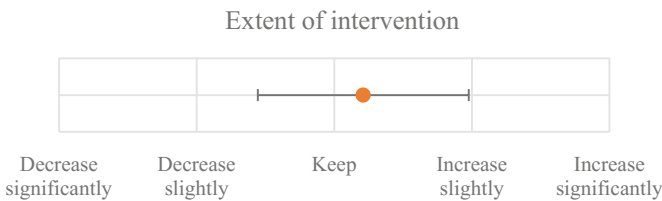


Fig. 7. Results of subjective ratings for extent of initiative for initial mission planning (mean and standard deviation, N = 20)

Results for subjective ratings for the agent intervention behavior during the initial planning can be seen in Fig. 6. The figure shows mean value and standard deviation. The figure shows subject’s ratings only for the 6 cases, when the agent generated a proposal. In general, these interventions during initial planning were rated rather unnecessary and retarding. This result confirms objective measured results.

6.2 Use Case: Re-planning Due to Threat

The second described use case comprises the change of the tactical situation due to a pop-up threat on the primary helicopter route in flight. The agent’s default mode in flight was the incremental planning mode. In such cases the agent immediately calculated an alternative route, which minimized the deviation from primary flight plan and proposed the solution to the pilot. Additionally, pilots could switch to the pre-planned alternative route (Fig. 2) or re-plan the route manually. Usually, subsequent to the re-planned helicopter route, a UAV had to be re-assigned to the new helicopter route for reconnaissance reasons. If the pilot did modify the UAV task within reasonable time (approximately 30 s after the first one), the agent generated a second proposal for this issue. Results show that 13 out of 18 proposals were accepted. Three times the pilot preferred a complete rerouting using the preplanned alternative route. In two cases, the pilot preferred a manual rerouting (Fig. 8). After half of the conducted missions, pilots were asked to rate the re-planning proposals generated by the planning agent. It can be seen that the interventions were accepted well in general (Fig. 9). The figure shows mean value and standard deviation. In a number of cases, subjects stated that the extent of intervention should be increased ($M = 0.40$, $SD = 0.70$) (Fig. 10).

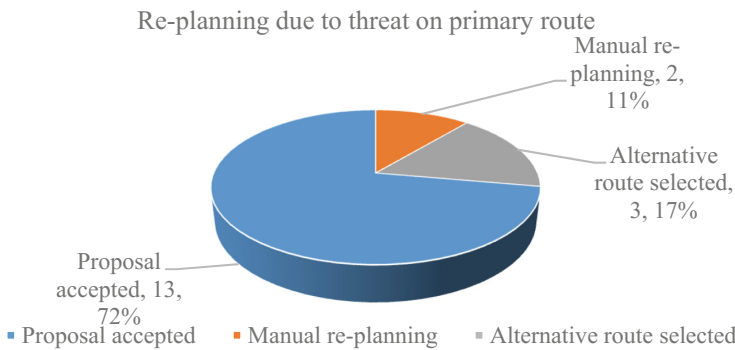


Fig. 8. Results of accepted and declined proposals during re-planning due to threat (N = 18)

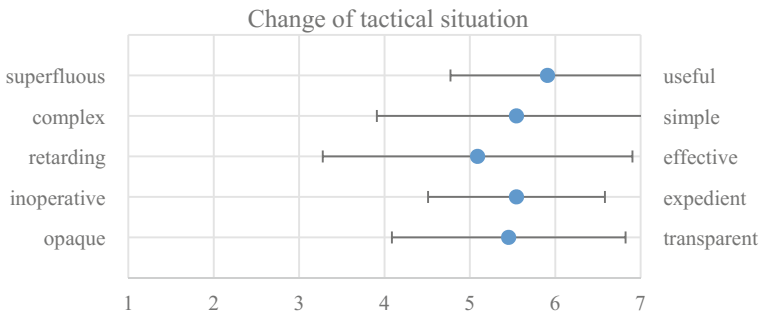


Fig. 9. Subjective ratings for agent behavior on threat detection on primary helicopter route (mean and standard deviation, N = 11)

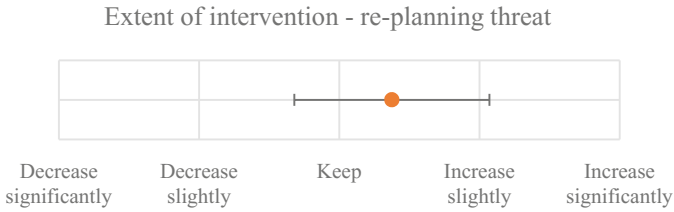


Fig. 10. Results of subjective ratings for extent of initiative for re-planning due to threat (mean and standard deviation, N = 11)

In these cases, the pilots mentioned that they would prefer the option to correct both, helicopter route and UAV task, using a single dialog.

6.3 Use Case: Re-planning Due to Change of Mission Objective

When the mission objective changed throughout a mission, the automation level of the planning agent was changed to the *extensive planning mode* due to an assumed workload peak. In this case, the agent generated a complex proposal comprising a helicopter route and a corresponding UAV task assignment for the two most important tasks (landing point detection and reconnaissance of primary route). On acceptance, the planning agent implemented these tasks automatically. Figure 11 shows the acceptance rates of proposals. In 19 use cases, (95%) the planning agent generated an extensive proposal. In 90% of all cases the proposal was accepted. The results show the high value of this type of intervention in the given situation.

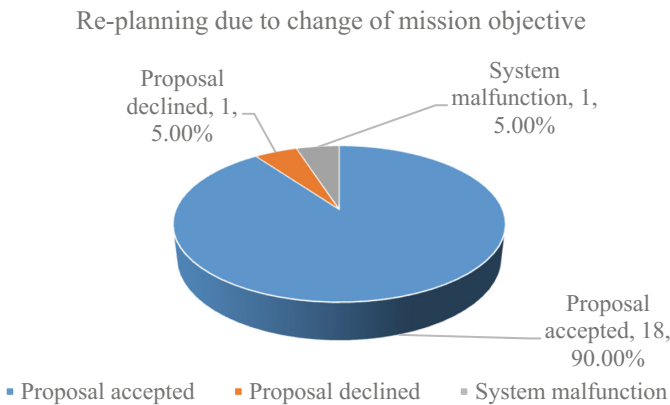


Fig. 11. Results of accepted and declined proposals immediately after change of mission objective (N = 20)

Figure 12 shows the subjective ratings for the agent behavior within the given re-planning use case. The overall ratings were very positive. The interventions were

rated helpful and effective. However, subjects also stated that the transparency of interventions could be increased. The extent of the re-planning proposals was rated very well ($M = 0.15$, $SD = 0.49$) (Fig. 13).

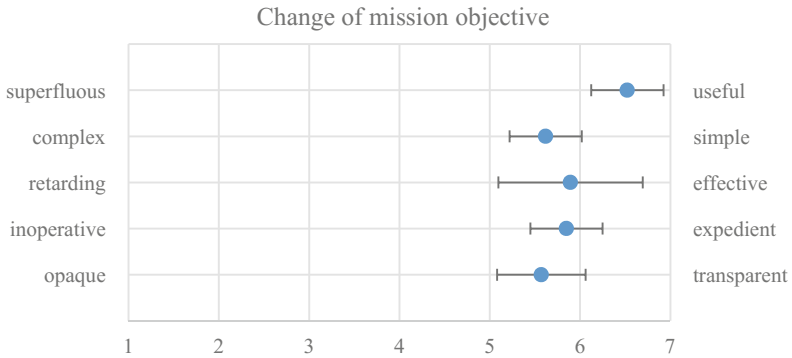


Fig. 12. Ratings for assistance after change of mission objective (mean and standard deviation, $N = 20$)

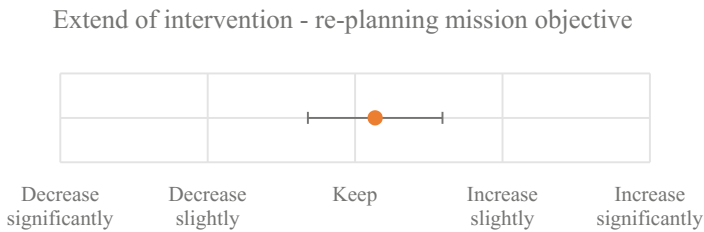


Fig. 13. Results of subjective ratings for extent of initiative for re-planning due to change of mission objective (mean and standard deviation, $N = 20$)

6.4 Overall Rating

After completion of the last mission, pilots had to rate the overall *mixed-initiate* agent behavior (Fig. 14). The figure shows that these ratings were very positive. Pilots stated that the behavior is expedient and useful. However, pilots also stated that transparency of the agent’s proposals could be increased. Several planning proposals were rather difficult to understand. This was especially the case, when the agent recommended an action, which was not visible to the pilot in the current map section. In these cases, pilots felt that they were interrupted in their work flow. Overall ratings comprise all interventions generated by the agent throughout the experimental campaign. Besides previously discussed types of intervention, this rating also includes agent task proposals for new UAV tasks.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|-------------|---|---|----|----|-----|----|----|-------------|
| superfluous | | | | | • | • | •• | useful |
| inoperative | | | | | • | •• | • | expedient |
| complex | | | • | | ••• | | | simple |
| ineffective | | | | • | • | •• | | effective |
| opaque | | | | • | •• | • | | transparent |
| slow | | | | •• | •• | | | fast |
| insistent | | | • | • | •• | | | restraint |
| as expected | | | •• | • | • | | | surprising |

Fig. 14. Overall ratings

7 Conclusion

In this article, we presented a concept for a *mixed-initiative* planning associate, which is designed to support the crew of a manned helicopter managing a MUM-T mission with up to three UAVs. Our scalable approach offered assistance on different levels of automation, depending on the tactical situation and the workload of the operator. Thereby, workload was determined by an external workload component.

We designed a full-mission experiment with crews of German army aviators. So far, we focused on the pilot’s subjective feedbacks regarding the behavior of the planning agent. Results indicate that the *scalable mixed-initiative* approach was accepted in general very well. When looking into the details, ratings could be categorized based on the three use cases. Agent interactions, generated during the initial mission planning on ground, were rated rather superfluous. Pilots would like to have either no proposals or an extensive planning proposal, which can be modified afterwards through the pilot. Agent interventions for re-planning support on changes of the tactical situation were received very well. A particularly high rating was noticeable for agent behavior after changes of the mission objective. The varying extent of intervention was received very well for all three use cases. This shows the huge advantage of our *scalability* concept in the *mixed-initiative* approach.

The evaluation of objective data obtained through the described experiment is ongoing. Future work will comprise *mixed-initiative* planning with multiple human users.

References

1. Strenzke, R., Schulte, A.: Design and evaluation of a system for mixed-initiative operation. *Acta Futura* 5, 83–97 (2012)
2. Clare, A.S., Macbeth, J.C., Cummings, M.L., Member, S.: Mixed-initiative strategies for real-time scheduling of multiple unmanned vehicles. In: American Control Conference, pp. 676–682 (2012)

3. Bresina, J.L., Jonsson, A.K., Morris, P.H., Rajan, K.: Mixed-initiative planning in MAPGEN: capabilities and shortcomings. In: Proceedings of the ICAPS–2005 Workshop on Mixed-Initiative Planning and Scheduling, p. 8 (2005)
4. Allen, J., Ferguson, G.: Human-machine collaborative planning. In: NASA Planning and Scheduling Work (2002)
5. Schulte, A., Donath, D.: A design and description method for human-autonomy teaming systems. In: Advances in Intelligent Systems and Computing, pp. 3–9 (2018)
6. Schmitt, F., Roth, G., Schulte, A.: Design and evaluation of a mixed-initiative planner for multi-vehicle missions. In: Harris, D. (ed.) EPCE 2017. LNCS (LNAI), vol. 10276, pp. 375–392. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58475-1_28
7. Simon, H.A.: Rational choice and the structure of the environment. *Psychol. Rev.* **63**, 129–138 (1956)
8. McDermott, D., Ghallab, M., Howe, A., Knoblock, C.: PDDL-the planning domain definition language. In: AIPS 1998 Plan (1998)
9. IBM Corp., IBM: V12. 1: User’s Manual for CPLEX. *Int. Bus. Mach. Corp.* **12**, 481 (2009)
10. Honecker, F., Brand, Y., Schulte, A.: A task-centered approach for workload-adaptive pilot associate systems. In: Proceedings of the 32rd Conference of the European Association for Aviation Psychology – Thinking High AND Low: Cognition and Decision Making in Aviation, Cascais, Portugal (2016)
11. Brand, Y., Schulte, A.: Model-based prediction of workload for adaptive associate systems. In: Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1722–1727 (2017)
12. Schmitt, F., Roth, G., Barber, D., Chen, J., Schulte, A.: Experimental validation of pilot situation awareness enhancement through transparency design of a scalable mixed-initiative mission planner. In: Karwowski, W., Ahram, T. (eds.) IHSI 2018. AISC, vol. 722, pp. 209–215. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73888-8_33



Flight Safety: ESL Flight Crew Member Use of Crew Alerting and Information Systems

Dujuan Sevillian^(✉)

Human Factors and Transport Systems Engineering,
Cranfield University, Bedford, UK
d.sevillian@cranfield.ac.uk

Abstract. Flight Deck Crew Alerting Systems (FDCAS)—alert systems/Quick Reference Handbook (QRH) checklists are designed with an understanding of how to effectively integrate human capabilities with alert system complexities. There are many English as-a-second language (ESL) flight crewmembers that utilize western designed FDCAS. Purpose of this study was to determine if ESL flight crewmembers' performance was impacted by use of western built FDCAS during non-normal conditions. Results indicated that ESL flight crewmember English language proficiency and background knowledge were factors that influenced their performance when they utilize crew alerting systems and QRH checklists during non-normal conditions. Design and integration of English language on crew alerting systems and QRH checklists were also contributory factors that impacted flight crewmembers' performance.

Keywords: Human factors · Flight safety · System safety
Flight deck systems · ESL · Lexis

1 Introduction

On the flight deck, English as-a-second language (ESL) flight crewmembers (captain and first officer) use crew-alerting systems designed with an English language emphasis. Design and integration of written English language on crew alerting and information systems (e.g. Quick Reference Handbook (QRH)) should provide ESL flight crewmembers with information enabling them to read and comprehend information adequately. Written English language is the preferred language of aviation (Hutchins et al. 2006) and it is utilized by ESL flight crewmembers to read and understand normal and non-normal conditions that may occur during typical phases of flight. Therefore, ESL flight crewmember ability to read and comprehend written English language and their level of English language proficiency should be adequate. Government and academia have investigated fundamental challenges ESL flight crewmembers experience while reading and comprehending written English language on the flight deck. Some of these challenges were discussed in multiple studies

Note: Following sections: (Methods, Results and Discussion) are from publication Sevillian (2017).

conducted by the Federal Aviation Administration (FAA). These studies noted that cryptic messages on crew alerting systems have the potential to impact ESL flight crewmembers' ability to read and comprehend information on displays (FAA 1996; 2013). It was also recommended that simplified technical English be used when designing crew-alerting systems. However, FAA studies did not elaborate on the impact written English language has (e.g. QRH information), on ESL flight crewmember performance. In other words, FAA studies should have provided evidence of written English language design and integration factors that effect ESL flight crewmembers' reading comprehension. Aircraft accident reports reveal that written English language has contributed to several airline accidents. In 2014 an A-320 aircraft flown by ESL flight crewmembers crashed. According to the investigation report, one factor that could have contributed to the accident was flight crewmembers' ability to read and understand written English language on a checklist. Flight crewmembers indicated they were challenged with information on the QRH checklist related to computer-reset functionality. The checklist confused flight crewmembers and did not provide them adequate understanding of the non-normal condition (KNKT 2015). In 2009, ESL flight crewmembers were involved in an A-330 aircraft accident. The accident report suggested that detailed written English language procedures may have negatively impacted flight crewmembers' performance (BEA 2012). The Center for Investigation and Prevention of Accidents (CENIPA) investigated an aircraft accident in year 2011 related to ESL flight crewmember misunderstanding of written English language on the QRH checklist. The accident report indicated that there were many checklist inaccuracies that negatively impacted ESL flight crewmembers' ability to read and comprehend information on the checklist (CENIPA 2013). Another aircraft accident involving a MD-83 in 2014 was related to ESL flight crewmembers inadequate response time and awareness to airspeed and written English language in the Flight Crew Operations Manual (FCOM). Information in the FCOM contained text related to anti-icing systems with ambiguous wording related to procedure protocol for responding to the aircraft condition (MCI 2016). In yet another aircraft accident involving ESL flight crewmembers, the investigation revealed that ESL flight crewmember written English language proficiency was a factor that influenced the accident. English as-a-second language flight crewmembers' ability to understand written English language on technical documentation impacted their ability to use information to solve problems on the flight deck. Their background knowledge of the written English information on Western built aircraft negatively impacted their reading comprehension performance (IAC 2009). Each of the previous accidents discussed reveal that English language proficiency, ESL flight crewmember background knowledge, and design/integration of written English language on information systems are factors that negatively impact ESL flight crewmember performance.

Although the aforesaid accidents reveal the outcome of ESL flight crewmember linguistic challenges on the flight deck, other studies reveal factors related to ESL flight crewmember linguistic challenges. The Aerospace Industries Association (AIA) conducted a study regarding ESL flight crewmember response to propulsion failures. The AIA indicated that approximately 15% of National Transportation Safety Board (NTSB) investigations were related to ESL flight crewmember ability to adapt to written English language. The study concluded that more emphasis on improving

written English language on propulsion system diagnostics is needed when utilized by ESL flight crewmembers (Sallie and Gibbons 1998). A university in China revealed that 80% of their aviation students have experienced various challenges related to written English language. Challenges were related to reading comprehension of written English language, with respect to vocabulary words on technical documentation (Wang 2011a). Likewise, Ho (1996) revealed similar issues in a study that focused on flight deck operations procedures manuals. In Ho's (1996) study, 30% of ESL flight crewmembers did not understand written English language safety data on documentation that referenced non-normal conditions. Smith-Jackson and Wogalter (2000), revealed that warnings should be understood by individuals with different linguistic backgrounds. Wogalter et al. (1997) indicated safety information related to warnings should be read and comprehended adequately by ESL adults, especially when they perform in sociotechnical environments (e.g. flight deck). Each of the previously mentioned studies provides indication that individuals with different language backgrounds should understand crew alerts. It was also recommended that more research is needed to understand impact of ESL flight crewmember understanding of system diagnostics that contain written English language emphasis.

Other studies have indicated that ESL flight crewmembers translate information on crew alerting systems and checklists. However, flight crewmembers should not have to translate information back in to their native language, given their proficiency reading and comprehending English language is adequate. If the meaning of vocabulary words/text corpora is cognitively translated by flight crewmembers in to their native language, it could impact their understanding of the crew alert and subsequent procedures (Drury and Ma 2005). If flight crewmembers translate written English language into their native language and they misunderstand vocabulary words and/or syntax meaning, this may cause flight crewmembers to revert back to their native language and search for words compatible to read and understand written English language (Kobayashi and Rinnert 1992). These types of factors have the potential to impact ESL flight crewmember performance and flight safety. It is obvious that precursors to ESL flight crewmember misunderstanding of written English language can negatively impact their performance. These precursors are mainly design and integration of written English language, ESL flight crewmembers' proficiency, background knowledge, and their performance related to use of written English language on the flight deck. Linguistic challenges that impact ESL flight crewmember performance on the flight deck need further investigation. Forthcoming literature provides more evidence of factors that influence ESL flight crewmember performance on the flight deck.

2 Literature Review

Discussion of previous aircraft accidents do not provide enough details to understand fundamental problems ESL flight crewmembers experienced on the flight deck. Government and academia have addressed symptoms leading to ESL flight crewmember/individual misunderstandings related to written English language. Following literature review provides an overview of many factors that influence ESL adult ability to read and comprehend written English language. Next studies provide a

review of fundamental challenges ESL adults experience while reading and comprehending written English language related to vocabulary words, text corpora, translation of text, design and integration of text, and English language proficiency. It also reviews methods that ESL adults use to read and comprehend written English language such as metacognition and use of background knowledge. These methods will be further discussed in the forthcoming review of literature.

According to Hancock (1998), reading requires processing and understanding information. Lexical knowledge and skills are acquired as a result of reading information. Comprehension requires an individual to understand what they are reading and utilize strategies (metacognition) to understand syntax and apply vocabulary knowledge. (Baker and Brown 1984). Cognition is defined as ability to read and comprehend information and apply it to a contextual environment. These factors require an individual to have a level of proficiency regarding written English language. English as-a-second language adults experience difficulties reading and comprehending syntax (e.g. sentences) (Condelli and Wrigley 2006). Karbalaei (2010) indicated that ESL adult ability to read and understand written English language is predicated by strategies they use to process their ideas, and this aids in execution of their decision-making processes. There are various types of mental models that ESL adults utilize to read and understand written English language. Top down and Bottom up models (or a combination of thereof) aid ESL adults with reading and understanding written English language. Bottom up models help ESL adults understand flow of linear text. This requires ESL adults to understand letters, vocabulary words, and phrases, and then decode the sentence meaning. This model is highly dependent on ESL adult level of written English language proficiency. Top down model consists of ESL adult use of background knowledge to understand information they read. Parry (1991) studied effect of mental model use by ESL adults while they read text. Small population of ESL adults utilized bottom up strategy while reading text corpora that was complex (contained challenging vocabulary words). English as-a-second language adult English language proficiency was low or intermediate level. Goal of the study was to determine if ESL adults could identify complex words and understand word meaning. Results indicated that ESL adults were challenged by vocabulary words and often omitted words. It was also indicated that ESL adults need extra time to interpret words that they did not know. It was revealed that due to their English language proficiency levels, their ability to understand text was negatively impacted. Yildiz-Genc's (2009) developed a study that focused on 15 ESL adults with intermediate English language proficiency. Adults read text without a time constraint and two strategies (bottom up and top down) were utilized to read text. It was noted that top down strategy was utilized by ESL adults due to their proficiency level. Adults indicated they were challenged with word meaning while reading text. They utilized sentence syntax to understand information (e.g. phrases) from previous sentences. They also utilized their background knowledge and metacognitive strategies to understand text. Use of background knowledge by an individual may be utilized to facilitate lexical inferring to understand word meaning. Parry (1991) and Yildiz-Genc's (2009) studies reveal English language proficiency is a key factor that influences ESL adult ability to interpret words. However, their studies do not focus on vocabulary word type and the effect ESL adult ability to process and interpret word meaning. Both studies indicate metacognitive strategies

are important methods ESL adults utilize to read and comprehend written English language, and some adults use these strategies based on background knowledge and English language proficiency.

Next studies review types of vocabulary words and how their characteristics impact ESL adult reading comprehension. These studies also reveal impact ESL adult background knowledge and written English language proficiency has on their reading comprehension performance. Wang (2011b) utilized 34 ESL adults with advanced English language proficiency. Wang's study researched use of lexical inference strategy use by ESL adults when they read an article with 240 vocabulary words. Each of the adults had experience (i.e. background knowledge) with academic and technical written English language. Results indicated that ESL adults utilized collocation knowledge and word association (e.g. lexical inferencing) to read and interpret written English language. Some of the participants incorrectly interpreted words they were not accustomed to reading in the text. Dycus (1997) indicated that highly proficient adults with vast vocabulary knowledge are often correct when they make inferences on vocabulary words. Adults with low English language proficiency and low vocabulary knowledge often make incorrect inferences on vocabulary words in text. In this case, highly proficient adults experienced difficulties interpreting types vocabulary words. Wang's (2011b) study does not corroborate Dycus (1997) finding that highly proficient adults should not make incorrect inferences on vocabulary words in text. However, in Wang's study these results could have also been due to adult level of background knowledge. Dwaik and Shehadeh (2013) researched the impact of lexical inferencing strategy had on 60 adults with low and high written English language proficiency levels. Results indicated that adults with high written English language proficiency guessed more words correctly than adults with low written English language proficiency. Adults with low written English language proficiency were challenged by use of context clues in sentences to read and understand text. Dwaik and Shehadeh (2013) and Dycus (1997) indicated that low English language proficiency adults and low vocabulary depth often experience difficulties with their interpretation of vocabulary words. Wang's (2011b) study revealed strategies are important factors that influence adult ability to interpret written English language and that adult English language proficiency varies and is influenced by strategy use and background knowledge.

Next studies provide an overview of how written English language design and integration has the potential to impact ESL adult reading comprehension. Simplified written English text and sentence length are factors that influence ESL adult reading comprehension performance. Mehrpour and Riazi (2004) utilized 100 adults that were proficient with their use of written English language. Half of the adults had a background in English language and the other half did not. Adults had approximately five years of English language experience. Each of the adults read technical and academic texts. The first text (i.e. medicine genre) had approximately 240 vocabulary words, while the other text (i.e. sociology genre) had 260 vocabulary words. Results indicated that shorter text was more difficult to read than longer lengths of text. Abdul-Hamid and Samuel (2012) researched the impact of ESL adult reading comprehension performance while reading two types of scientific texts and metacognitive strategy used. Adult English language proficiency was proficient or not proficient. Goal of the research study was to determine level of difficulty when ESL adults read the texts. In

the first text, participants had background knowledge of half the text they read in their native language. English as-a-second language adults were familiar with 30% of the second text. First text contained approximately 590 vocabulary words, while the other text contained approximately 740 words. Results indicated long sentences in both text were difficult to read and led to re-reading text. Adult level of proficiency was a factor influencing their ability to read and comprehend each of the scientific texts. It was also indicated that ESL adults translated vocabulary words into their native language to understand texts. Kim (2006) conducted a study on the impact of ESL adult reading comprehension when they read text with abbreviations/acronyms. Adult English language proficiency was low and high levels. They had three years of English language background knowledge and translating written English language text into their native language was a common strategy used to read and comprehend English language. Results indicated that acronyms/abbreviations were difficult for ESL adults to read and interpret because of different word meanings that are commonly used in ESL adult native language. You (2009) investigated the impact of ESL adult ability to read information on computer screens and on paper format. There were 120 ESL adults that participated in the study that had background knowledge in using English language. Two texts were utilized for the experiment. One text was familiar to the adults, while the other text was not. Adult English language proficiency was high, medium, and low. Each text contained 340 words. Paper format text allowed for more space for lines of text versus the computer screen format, which allowed for less text. Results indicated that adults performed better while reading text from the paper than from the computer screen. This was likely due to ESL adult ability to highlight information on paper and use other strategies (e.g. re-reading text) to read written English language on paper format. Adults did not use many strategies to read information on computers screens. But, adults indicated they were comfortable with their background knowledge while reading and comprehending information on the computer screen. Adults with low English language proficiency level reading performance was not adequate when they read information from paper and computer screen formats. On the other hand, medium and high proficiency level adults performed well when they read and interpreted information in each of the text formats. Mehrpour and Riazi (2004) and Abdul-Hamid and Samuel (2012) studies revealed that the type of text adults read, English language proficiency level, level of background knowledge, strategy use, and short versus long strings of text influences adult reading comprehension performance. Kim (2006) and You (2009) indicated that acronyms and abbreviations have an impact on adult performance depending on their background knowledge of the long form (e.g. HYD versus Hydraulic). Both authors reveal that information on a screen versus information on paper has an impact on adult performance. This is due to type of metacognitive strategy utilized by adults, their English language proficiency, and background knowledge. Although vocabulary word type, text type, adult English language proficiency, background knowledge, and strategy use are factors that influence adult performance while reading and comprehending written English text, there are adults that cognitively translate information into their native language to have a better understanding of written English language. On the other hand, translation of text by a translator also reveals important details as well.

Forthcoming studies review translation of text and impact on ESL adult reading comprehension performance.

Translation of written English language can occur in two different ways: unilateral translation of written English language text or translation of written English language text into ESL native language by a translator). Each type of translation has the potential to impact ESL adult ability to read and understand written English language. According to Ogilvie (1984), due to various complexities in written English language, it may not be appropriate to translate text into an adults' native language. Zhao (2015) conducted a research study with 15 ESL adults, and investigated the impact of translating written English language into adult native lexis. Results indicated that adults that have adequate background knowledge of their native language are better equipped to read and comprehend a second language. It was also indicated that ESL adult inability to comprehend vocabulary words that were translated was due to adult ability to understand and translate word meaning into ESL adult native language. Ynfiesta et al. (2013) developed a study to determine impact of translating written English language acronyms used in technical text into a different another language. An experienced translator performed the translation task. It was determined that background knowledge of the long form acronym in written English language is essential, so that there are no misunderstandings of word meaning in ESL adult native language. The translator often searched for words that had equivalent meaning in another native language. It was noted that aforesaid factors have the potential to impact on ESL adult reading comprehension performance. Barani and Karimnia (2014) studied the impact of written English language translated into 32 ESL adults' native language (e.g. Persian language). Goal of the study was determine strategies used to read text that was translated from English language into their native language. Results indicated participants' background knowledge in the text genre (i.e. scientific text) enabled them to understand text. They used several metacognitive strategies (i.e. unilateral translation) to understand the text that was translated from written English language into Persian lexis. Zhao (2015) and Ynfiesta et al. (2013) indicated that translation of written English language words into adult native language impact reading comprehension performance. Zhao (2015) provided evidence that background knowledge of vocabulary words in adult native language can help them understand a second language. However, in each study adult English language proficiency was not reviewed. In previous studies, English language proficiency was noted as a factor that impacts adult understanding of vocabulary words and their meanings. In Yinefista et al. (2013) study it was revealed that information translated by a translator can impact adult performance. This is due to background knowledge of the translator and his/her ability to connect the translated word meaning to ensure that the reader understood it. Barani and Karimnia (2014) corroborated Zhao (2015) and Ynfiesta et al. (2013) studies regarding the need for background knowledge to understand text translated from written English language text to adult native language.

Overall, the literature review provides evidence of factors that influence ESL adult ability to understand written English language. Vocabulary words, text genre, adult English language proficiency, background knowledge, and strategy use impact adult performance. Unilateral cognitive translation of written English language into adult native language reveals challenges. Likewise, having a translator translate information

from written English language into adult native language requires background knowledge of different types of text, vocabulary words and their meanings. In the context of ESL flight crewmembers use of procedures and crew alerts utilization on the flight deck, do flight crewmembers experience performance challenges while reading and comprehending written English language on the flight deck? Do ESL flight crewmembers' linguistic challenges impact flight safety? It is hypothesized that there will be a statistically significant difference and interaction between ESL flight crewmember reading comprehension proficiency and performance when they read and comprehend written English language on QRH checklists and ECAM system, and written English language on QRH checklists translated into ESL flight crewmembers' native language.

3 Methods

Thirty male ESL flight crewmembers from Portugal with air transport pilot ratings that currently fly aircraft for an airline were utilized for the study. Each flight crewmember had experience flying several Airbus aircraft types (i.e. A319). Flight crewmembers' native language was Portuguese. Flight crewmembers English language was learned through formal schooling (i.e. high school) and this was considered background knowledge. Flight crewmembers' International Civil Aviation Organization (ICAO) English language proficiency levels were utilized as their background knowledge using English language. Flight crewmembers ICAO English language Proficiency Rating (ELPRs) met the minimum level four operational. Level four operational indicates that flight crewmembers have adequate use of English language (e.g. speaking/listening). The ELPRs provide the reader with an understanding of flight crewmembers background knowledge of English language. Flight crewmembers rated their Reading Comprehension Level (RCL). Ratings were utilized to describe flight crewmembers' proficiency ratings. Questionnaires were provided to each flight crewmember asking them to rate their proficiency, when they read and comprehend written English language on crew alerting systems and QRH checklists. Proficiency ratings were determined to be high level or medium level. High-level proficiency rating indicated flight crewmembers understood written English language vocabulary words, while medium-level proficiency rating indicated flight crewmembers experienced difficulties with use of vocabulary words/word meaning. Proficiency levels were utilized to determine flight crewmember extent of reading and interpreting English language, and if differences exist between flight crewmembers proficiency levels. A within subjects experimental design was developed and contained independent variable (IV)-language and dependent variables (DV)-response time, and National Aeronautics Space Administration (NASA) Task Load Index (TLX) workload scores. English as-a-second language flight crewmember response time was measured with a stopwatch, starting with the outset of the alert and time was stopped when the trial was complete. Electrical and hydraulic alerts from the A-320 aircraft Electronic Centralized Aircraft Monitor (ECAM) were utilized in the experiment. The NASA TLX workload scores were recorded post task completion for each system fault. Experimental trials lasted for sixty minutes. Each flight crewmember piloted an A-320 flight deck for 30 min while the

researcher injected faults (electrical and hydraulic) during cruise phase of flight. Last thirty minutes was allocated for post interview discussion with each flight crewmember. Prior to the start of the trials, the researcher evaluated written English language text on written English language ECAM and QRH checklists, to determine text genre and vocabulary word types. Written English language on the QRH checklists were translated into Portuguese language by translators at the airline. More details on the translation method will be provided in a forthcoming section.

Limitations. Written English language vocabulary words and text genre from electrical and hydraulic ECAM and QRH checklists were utilized for the study. Flight crewmember use of different written English language on ECAM and QRH checklists (e.g. pneumatic system) may have impacted their performance differently. Translation of QRH checklists into a different lexis (e.g. Chinese) may have impacted flight crewmember performance. Flight crewmember background knowledge of information, English language proficiency, use of metacognitive strategies may have also impacted their performance while using different crew alerting and information systems.

Table 1 is a review four specific hypotheses. The format of the hypotheses is as follows: Hypothesis, condition, and null hypothesis.

Text corpora on ECAM and QRH checklists were evaluated prior to the experimental trials. As the literature review indicated, several vocabulary word types and text genre can be found in text corpora. An evaluation of text prior to flight crewmembers participation in the experimental trials was conducted. Translators, with experience in translation methods translated written English language QRH checklists (i.e. hydraulic and electrical system) into Portuguese language. Abbreviations, phrases, and acronyms were not translated if there was no equivalent meaning in Portuguese language. Previously mentioned, it is important to be aware that translation can impact adult understanding of word meaning and cause interpretation issues. Written English language ECAM system and QRH checklists text were not altered. In other words, authentic text was utilized for the study, certified by the airline. Texts were not simplified, word tokens were unchanged, and sentence length was not manipulated. If the ECAM system and QRH checklists had been altered prior to the study, results may be different. Fonts, and word case tense was unchanged from its original format. Text genre on the ECAM system and QRH checklists contained technical information with several different vocabulary word types. Furthermore, text contained expository and instructional information. Researcher utilized authority references such General Service List of English Words (GSLEW), Academic Word List (AWL), and the A-320 Flight Crew Training Manual (FCTM), ECAM system manual to evaluate written English language ECAM texts and QRH checklists texts. Some of the authority references contained technical/scientific, sub-technical, non-technical, and acronyms/abbreviations/long form word types, which were also found on the ECAM and QRH checklists. Each word on the ECAM and QRH checklists was mapped to the authority references. Results indicated a high percentage of high frequency words and several occurrences of words from the AWL and GSLEW lists. There were a small number of low frequency words and many sub-technical and technical words found on the ECAM and QRH checklists. Since written English language (in general) contains many words found on GSLEW and AWL lists Coxhead (1998) and West (1953), participants with a background in written

Table 1. Listed and described hypotheses tested

| Hypothesis #1 (H_A) | Condition | Null Hypothesis #1 (H₀) |
|--|--|--|
| There will be a significant difference between participant performance with use of ECAM (written English language)/written English language QRH checklists and ECAM (written English language)/Portuguese language QRH checklists, and participant response time to electrical and hydraulic system malfunctions. | Participant response time will be slow with use of ECAM (written English language)/written English language QRH checklists and fast with use of ECAM (written English language)/written QRH checklists Portuguese language when participants respond to electrical and hydraulic system malfunctions. | There will not be a significant difference between participant performance with use of ECAM (written English language)/QRH checklists and ECAM (written English language)/written Portuguese language QRH checklists, and participant response time to electrical and hydraulic system malfunctions. |
| Hypothesis #2 (H_A) | Condition | Null Hypothesis (H₀) |
| There will be a significant difference between participant performance with use of ECAM (written English language)/written English language QRH checklists and their NASA Task Loading Index (TLX) workload scores, and when they use the ECAM (written English language)/written Portuguese language QRH checklists and their NASA TLX workload scores. | Participant NASA TLX workload scores will be high with use of ECAM (written English language)/written English language QRH checklists, and participant NASA TLX workload scores will be low with use of ECAM (written English language)/written Portuguese language QRH checklists, when participants respond to electrical and hydraulic system malfunctions. | There will not be a significant difference between participant performance with use of ECAM (written English language)/written English language QRH checklists and ECAM (written English language)/written Portuguese language QRH checklists and participant NASA TLX workload scores, when they respond to electrical and hydraulic system malfunctions. |
| Hypothesis #3 (H_A) | Condition | Null Hypothesis (H₀) |
| There will be a significant positive correlation between participant NASA TLX workload scores (ECAM written English language/written Portuguese language QRH checklists) and participant response time (ECAM written English language/written Portuguese language QRH checklists) | As participants' NASA TLX workload scores decrease while using ECAM written English language/written Portuguese language QRH checklists so will their response time using ECAM written English language/written Portuguese language QRH checklists | There will not be a significant positive correlation between participant NASA TLX workload scores (written English language ECAM)/(written English language/written Portuguese language QRH checklists) and participant response time (ECAM written English language/written Portuguese language QRH checklists) |
| Hypothesis #4 (H_A) | Condition | Null Hypothesis (H₀) |
| There will be a significant positive correlation between participant NASA TLX workload scores and their use of written English language ECAM/written English language QRH checklists, and their written English language ECAM/written English language QRH checklists response times. | As participants' NASA TLX workload scores increase while using ECAM written English language/written English language QRH checklists, so will their response time using written English language ECAM/written English language QRH checklists. | There will not be a significant positive correlation between participant NASA TLX workload scores and their use of written English language ECAM/written English language QRH checklists, and their written English language ECAM/written English language QRH checklists response times. |

English language and adequate English language proficiency may benefit from such words found on the ECAM and QRH checklists. Regarding technical words/acronyms/phrases, there were many of these types of words on each of texts. As Coady and Huckin (1997), Chung and Nation (2004) indicated, technical words are required to be known by ESL adults based on their training and background knowledge of the technical field. Technical vocabulary has the potential to cause difficulties with ESL adult interpretation when reading text that is considered technical. It was also indicated that their proficiency is a key factor that influences their ability to read and interpret technical information. Regarding text layout, ECAM and QRH checklists had different layouts with respect to data presentation. As previously stated, abbreviations, acronyms, and phrases appeared differently in format, with respect to ECAM and QRH checklists. As indicated by

Hartley (1994), abbreviations and acronyms should be designed adequately so that technical information on checklists may be followed by ESL flight crewmembers, and thus allowing them to respond effectively to an alert. According to Dyson (2004), configuration of data may impact reading comprehension of information on paper. Configuration of data can also impact ESL flight crewmembers information processing on displayed crew alerts. Inter-rater reliability analyses were conducted to ensure there was no bias with categorizing the previously mentioned vocabulary words on the written English language ECAM and QRH checklists. Cohen’s Kappa coefficient was $k = 0.57$ for the ECAM electrical system and $k = 1$ for the electrical system QRH checklist. Cohen’s Kappa coefficient was $k = .55$ ECAM hydraulic system and $k = 1$ QRH hydraulic system checklist.

As the literature review indicated, it is essential to follow a methodical approach when translating information from one language to another. Authentic written English language selected QRH checklists (electrical and hydraulics) were translated from written English language into Portuguese lexis. Translation process lasted for one week and was conducted with two experienced translators. Both translators were ESL senior airline flight crewmembers whose first language was Portuguese. Each of the flight crewmembers rated their English language proficiency as high level. Following 14-step process was utilized to translate the texts (Table 2).

Table 2. Translation process

| | |
|---|--|
| 1.) Ensure texts are authentic and unchanged from original format. | 2.) Determination of translatable and non-translatable technical information items. |
| 3.) Non-translatable technical information- any written English language on the QRH checklist that corresponds to participant inputs on flight deck crew alerting systems or its interfaces (labels/panels/buttons/switches) that are written English language acronyms, abbreviations, or phrases with no equivalent meaning in Portuguese language. | 4.) Translatable technical documentation- Information associated with QRH checklist, notes—which included abbreviations, acronyms, and phrases, with equivalent meaning in Portuguese language. Or, non-flight deck input related information such as non-system command inputs by the pilot (i.e. sentences related to safety assurance, or reminders, phrases, notable information, with equivalent meaning in Portuguese language). |
| 5.) Review of aircraft technical illustrations (flight deck overhead panel and other related interfaces). | 6.) Matching exercise between QRH checklist technical information and flight deck technical information illustrations, to determine participant best mapping between flight deck crew alerting system interfaces and QRH checklist items. |
| 7.) Review of technical and non-technical items with association representative/senior pilot. | 8.) Preliminary review of QRH checklist translation process considered the country’s regional pedagogical approaches to teaching Portuguese language in Lisbon, Portugal. This review was needed to understand how participants’ read and comprehend Portuguese language when using technical information on the flight deck. |
| 9.) Syntax Exercise and Translation: Arrangement of words, acronyms, abbreviations, phrases, and sentences on checklist. Written English language technical information was not translated into Portuguese language if there was no equivalent word meaning in Portuguese language. | 10.) Assurance of font, color, and sentence spacing accuracy was conducted by ensuring written English language checklist font colors and character sizing was the same on the translated checklist. |
| 11.) Review of translation by association representative, senior pilot, and researcher for concurrence. | 12.) Printed copies of checklists (A4 paper 1 sided) 12.) Participants executed use of QRH checklists during experimental trials. |
| 13.) Participants executed use of QRH checklists during experimental trials. | 14.) Obtained verbal feedback regarding checklist design by participants after the trials. |

4 Results

Descriptive statistics indicated that the average age of flight crewmembers was 47 years and the minimum age was 27 years. Flight crewmembers' average airline years of experience was 24 years. Paired samples correlation test indicated mean response times from the written English language ECAM/written English language QRH checklists score was faster ($M = 8.75$; $SD = 3.811$) than participant response time on the Portuguese checklists ($M = 14.4$; $SD = 4.730$). The paired samples correlation value indicated a negative correlation ($-.075$), inverse relationship between participant response times when they utilize written English language ECAM/written English language QRH checklist and written English language ECAM/written Portuguese language QRH checklist. In other words, when participants use written English language ECAM/written Portuguese language QRH checklist to respond to hydraulic and electrical system malfunctions, they tend to have longer response times than with use of written English language ECAM/written English language QRH checklists. Significance value was ($Sig\ p = .695$). Since $p > .05$, this is an indication of no significant correlation. Paired samples t-test found a significant difference between participant response times when they use written English language ECAM/written English language QRH checklists and written English language ECAM/written Portuguese language QRH checklists. The results indicated $t(29) = -4.947$; $Sig\ 1\text{-tailed}\ p = 0$ and $Sig\ 2\text{-tailed}\ p = .000$; $p < .05$, $d = -.132$ (means are insufficient), the researcher accepts the alternative hypothesis (H_A) that there is a significant difference between participant response times when they use written English language ECAM/written English language QRH checklists, and written English language ECAM/written Portuguese language QRH checklists when participants respond to electrical and hydraulic system malfunctions. Participant response times with use of written English language ECAM/written Portuguese language QRH checklists was slow and their response time using written English ECAM/written English language QRH checklists was fast. A paired samples correlation was performed to determine if there would be a correlation between participants NASA TLX workload scores when they utilize the written English language ECAM/written English language QRH checklists/written English language ECAM/written Portuguese language QRH checklists. Results indicated that mean participant NASA TLX workload score from the written English language ECAM/written English language QRH checklists score was ($M = 34$; $SD = 17.777$), which was lower than participants NASA TLX workload score on the Portuguese checklists ($M = 50$; $SD = 23.163$). The correlation value was $.362$, indicating a positive correlation between the two variables (English language/Portuguese language). This is an indication that when participants utilized written ECAM written English language/written English language QRH checklists their NASA TLX workload scores tend to move in a positive direction, and when participants utilized ECAM written English language/written Portuguese language QRH checklists their NASA TLX workload scores tends to move in the positive direction. The paired samples correlation test indicated a significant correlation between participant NASA TLX workload scores when they use written English language ECAM/written English language QRH checklist and written English language ECAM/written Portuguese language QRH

checklist. The significance value for this analysis was $p = .049$, ($p < .05$) and the means are insufficient. This is an indication that there is a significant relationship between the aforesaid variables (English language/Portuguese language). Regarding the paired samples t-test, the researcher performed a one-tailed and two-tailed test and found a significant difference (both tests) between participant NASA TLX workload scores when they use written English language ECAM/written English language QRH checklists, and their NASA TLX workload scores when they use written English language ECAM/written Portuguese language QRH checklists. The values are as follows: $t(29) = -3.803$, (Sig. 1-tailed = .0005; 2-tailed $p = .001$) ($p < .05$), $d = -0.78$. Therefore, researcher accepts the alternative hypothesis (H_A) that there is a significant difference between participant written English language workload scores and Portuguese language workload scores, when participants respond to electrical and hydraulic system malfunctions. Participant use of written English language ECAM/written Portuguese language QRH checklists was more difficult than using written English ECAM/written English language QRH checklists. A Pearson product moment (Pearson's r) correlation test was performed to determine if a significant positive correlation exists between participant NASA TLX workload scores (ECAM written English language/written Portuguese language QRH checklists) and participant response time (ECAM written English language/written Portuguese language QRH checklists). Recall, participant NASA TLX workload scores were ($M = 50$; $SD = 23.163$) and response time was ($M = 14$; $SD = 4.730$) (higher workload scores and response times were observed when participants utilized ECAM written English language/Portuguese language QRH checklists, compared to their use of ECAM written English language/English language QRH checklists). Pearson correlation value was $r = .158$ which indicates a minimal positive correlation. This result indicates as participant NASA TLX workload scores increase so does their response time to hydraulic and electrical system malfunctions. The significance value was $p = .404$ ($p > .05$), $d = 2.15$. These results indicated no significant correlation between participant NASA TLX workload scores (ECAM written English language/written Portuguese language QRH checklists) and participant response time (ECAM written English language/written Portuguese language QRH checklists). The evidence suggests that the correlation observed is not generalizable to the population of ESL flight crewmembers. The researcher accepts the null hypothesis (H_0) that no significant positive correlation exists between participant NASA TLX workload scores (ECAM written English language/written Portuguese language QRH checklists) and participant response time (ECAM written English language/written Portuguese language QRH checklists).

A Pearson product moment (Pearson's r) correlation test was performed to determine if a correlation exists between participant use of ECAM written English language/written English language QRH checklists and their NASA TLX workload scores, and their use of ECAM/written English language/written English language QRH checklists response times. Recall, participant ECAM written English language/written English language QRH checklists NASA TLX workload scores mean was ($M = 34$; $SD = 17.777$) and ECAM written English language/written QRH checklists response times was ($M = 8.75$; $SD = 3.811$) (lower workload and lower response time observed when participant utilized ECAM written English language/written English language QRH checklists,

compared to their use of ECAM written English language/Portuguese language QRH checklists). The Pearson correlation value was $r = .150$ which indicates a minimal positive correlation. This result indicates as participant NASA TLX workload scores decrease so does their response time to electrical and hydraulic system faults. However, the significance value was $p = .428$ ($p > .05$), $d = 1.96$. These results indicated no significant positive correlation between participant NASA TLX workload scores (ECAM written English language/written English language QRH checklists) and participant response time (ECAM written English language/written English language QRH checklists). The evidence suggests that the correlation observed is not generalizable to the population of ESL flight crewmembers. The researcher accepts the null hypothesis (H_0) that no significant positive correlation exists between participant NASA TLX workload scores (ECAM written English language/written English language QRH checklists) and participant response time (ECAM written English language/written English language QRH checklists). Researcher developed hypotheses and corresponding two-way ANOVAs (between- subjects design) to determine effect of participant English language proficiency, airline years of experience, and impact on their reaction time/ NASA TLX workload scores (Table 3).

Table 3. Two-way ANOVAs between subjects hypotheses

| | |
|---|---|
| <p>H_A: There will be a significant main effect and interaction between participant airline years of experience/English language proficiency and their reaction time when they read and comprehend written English language on the ECAM/QRH checklists.</p> | <p>H_0: There will not be a significant main effect and interaction between participant airline years of experience/English language proficiency and their reaction time when they read and comprehend written English language on the ECAM/QRH checklists.</p> |
| <p>H_A: There will be a significant main effect and interaction between participant airline years of experience/English language proficiency and their NASA TLX workload scores when they read and comprehend written English language on the ECAM/QRH checklists.</p> | <p>H_0: There will not be a significant main effect and interaction between participant airline years of experience/English language proficiency and their NASA TLX workload scores when they read and comprehend written English language on the ECAM/QRH checklists.</p> |

No significant main effect and interaction were observed between participant airline experience, proficiency, and reaction time when they read and comprehend the written English language on crew alerting systems and QRH checklists. Results indicated $F(1, 26) = .003$, $p > .05$, partial $\eta^2 = .000$. Participant airline experience less than 20 years, high level proficiency participants reaction time mean was $M = 7.63$; $SD = 2.26$. Participant reaction time mean for medium level proficiency participants was $M = 9.00$; $SD = 0$. Participants with high-level proficiency reaction time were faster than medium level proficiency participants. Results also indicated $F(1, 26) = .046$, $p > .05$; partial $\eta^2 = .002$. Participant airline years of experience 20 years or greater and high level proficiency revealed their reaction time was $M = 9.62$; $SD = 4.66$.

Participants with medium level proficiency indicated $M = 7.25$; $SD = 1.32$. Participants with high-level proficiency had a longer reaction time than participants with medium proficiency level. Researcher accepts the null hypothesis. No significant main effect and interaction were observed between participant years of experience, proficiency, and NASA TLX workload scores when they read and comprehend written English language on crew alerting systems and QRH checklists. Results indicated $F(1, 26) = .028$, $p > .05$, $\text{partial } \eta^2 = .001$. Participants with less than 20 years of experience high level proficiency NASA TLX workload scores indicated $M = 40.26$; $SD = 18.96$. Medium level proficiency participants NASA TLX workload scores were $M = 15.00$; $SD = 0$. Participants with less than 20 airline years of experience high-level proficiency had higher NASA TLX workload scores than medium level proficiency participants. Results also indicated $F(1, 26) = 2.86$, $p > .05$; $\eta^2 = .099$. Participant airline experience 20 years or greater and high level proficiency indicated their NASA TLX workload scores $M = 34.66$; $SD = 17.21$. Participants medium level proficiency participants, $M = 24.15$; $SD = 16.7$. High-level proficiency participants with 20 years of experience or greater had higher workload scores than participants with medium level proficiency. The researcher accepts the null hypothesis.

5 Discussion

With respect to written English language on the ECAM and QRH checklists, participants' mean response times revealed they responded more quickly to electrical and hydraulic system faults than when they utilized English language translated into Portuguese language on QRH checklists. All participants had background knowledge reading and interpreting written English language. They also had experience with use of technical information on the flight deck while responding to non-normal conditions (i.e. system faults). Participants had experience with reading and comprehending information on different ECAM systems and QRH checklists. This enabled them to have an understanding of how written English language text was designed and integrated on the ECAM and QRH checklists. Participants indicated they responded quickly to alerts and use of written English language checklist because they were accustomed to the English language. It was noted that participants are trained on how to use technical information while responding to a system fault. Many of them indicated they have encountered non-normal conditions while flying aircrafts at their airline, and they are trained to understand written English language logic on ECAM and QRH checklists to ensure their response time is effective. During experimental trials, the researcher observed most of the participants responding to the system faults very quickly and with precision, with respect to following published QRH checklist procedures. Moreover, participants did not indicate issues with their use of written English language on the ECAM system. Technical information on the ECAM system and QRH checklists (abbreviations and acronyms) were familiar to many participants. Park's et al. (2014) study revealed that less time is utilized to read and comprehend acronyms, if ESL adults have sufficient amount of background knowledge of the acronyms in text. If longer response times are needed to process information such as acronyms/abbreviations on a display, it could impact their ability to solve time critical

system/aircraft problems. As the researcher did not regulate a time limit to complete each task, this could also be a reason that participant response time was fast when they responded to electrical and hydraulic system faults. Park's et al. (2014) study also provided an indication that temporal demand on ESL adults was not regulated when they read written English language text. Regarding participant English language proficiency and metacognitive strategy use in the researcher's study, participants had high and medium levels of English language proficiency and they used QRH checklist references (published FCOM procedure text) to assist them with responding to electrical and hydraulic system faults. As Park's et al. (2014) revealed, metacognitive strategy use such as referencing other sources is typical of ESL adults that have high level of English language proficiency. The researcher's findings support Park's et al. (2014) study. It was noted in the profiling of text exercise, there were many high frequency words (GSLEW) as well as academic words (AWL), small number of low frequency words, and sub technical/scientific acronyms/abbreviations. Previously discussed, written English language contains many high frequency words and they are more comprehensible due to their frequency in text (Nation 2001). Academic words were developed to catalog most frequently occurring words in academic text, and they assist learners of English-a-second language, with respect to their reading comprehension (Coxhead 1998). As participants had background knowledge of English language through different types of instructional learning, this could have prepared them for reading and understanding written English language. It should also be noted that the participants received written English language training in classes where there were different pedagogical approaches to teaching English language. This could also be a factor that influenced their ability to read and understand the language. Researcher's findings support Wanpen's et al. (2013) study, which indicated that taking courses in an English language curriculum helps facilitate reading comprehension of written English language. Participants also noted that since they were accustomed to written English language, they were able to use various strategies like decoding words, and re-reading words to help them through the reading comprehension process. Researcher's findings support Dwaik and Shehadeh (2013) and Nylander's (2014) studies, with respect to decoding vocabulary words (lexical inferencing) and participant English language proficiency.

Participants indicated they did not have background knowledge of written English language text translated into Portuguese language on QRH checklists. Participants indicated they often unilaterally translate vocabulary words into their native language (Portuguese), and that translation process occurs mostly under non-normal conditions. But, they do not translate every word on QRH checklists. It was noted, that translation processes occur if they have background knowledge of the English language vocabulary word/sentence in Portuguese language. As the airline indicated, it receives published/certified QRH checklists from the manufacturer that do not contain any changes to text. Portuguese flight crewmembers also indicated they are trained on text that appears on QRH checklists, which is provided to them by the manufacturer. Regarding participant's response time when they utilized Portuguese language on QRH checklists, their response time was slow. This could be due to participants' lack of background knowledge of translated text, and it could be that, they were aware of particular vocabulary words that had the same meaning in Portuguese language.

Participants indicated they re-read text due to uncertainties with word meaning in the translated text, monitored their reading speed due to their desire to make correct inferences on each word/sentence, and decoded words such as abbreviations/acronyms and other vocabulary words in the text. On the other hand, there were participants that read and comprehended Portuguese language text with ease, as they were familiar with text translated into Portuguese language that had an equivalent meaning. It was noted that aforesaid strategies used to read and comprehend Portuguese language slowed their response time to electrical and hydraulic system faults. On the contrary, they were comfortable with the time they spent reading and comprehending text, so that they would not make incorrect inputs on the flight deck. They were concerned if they read the text too fast, they would miss a word or omit information, which could also lead to long response times. Hutchins et al. (2006) and Drury and Ma (2005) indicated that translation of written English language has the potential to impact ESL adults reading comprehension. It was also noted by Al-Sohbani and Muthanna (2013) that participants must have background knowledge of written English language, so that they may adequately understand translated language. They must also have adequate English language proficiency. As most participants indicated, they had background knowledge of abbreviations/acronyms on crew alerting systems and QRH checklists. There were some participants that indicated abbreviations/acronyms long form was difficult to understand in English language. This could have negatively impacted their ability to understand English language translated into Portuguese language on QRH checklists. In Al-Sohbani and Muthanna (2013) study, participants did not have adequate knowledge of written English acronyms and abbreviations, and when acronyms and abbreviations were translated into their native language, they were difficult to read and understand word meaning. In the researcher's study, participants had adequate background knowledge and adequate English language proficiency when they use of written English language on crew alerting systems and QRH checklists. It is peculiar as to why their response time was longer on the written Portuguese language checklists than when they read and comprehended information on written English language crew alerting systems/QRH checklists. Throughout the researcher's experiment, participants often utilized metacognitive strategies to read and interpret Portuguese language (i.e. re-read sentences). They cognitively translated (unilaterally) Portuguese language into different vocabulary words to attain word meaning, and they also reverted back to using written English language. When participants re-translated Portuguese language text to attain other forms of vocabulary words in Portuguese language, this was most likely due to their misunderstandings of sentence syntax. They also reverted back to use of cognitive mental model of written English language on QRH checklists. According to Kobayashi and Rinnert (1992), reverting back to English language can occur because an ESL adult lacks understanding of translated syntax meaning. This behavior by ESL individuals can result in inappropriate translation of technical information back into their native language. Evidence from Barani and Karimnia's (2014) study suggested that many participants used metacognitive strategies such as re-read sentences and paraphrase words while they read English language text. It was indicated that they utilized these strategies for problem solving purposes, which were related to difficulties understanding word meaning. Part of Barani and Karimnia's (2014) study was corroborated in the researcher's study. The researcher found that participants re-read sentences to

understand word meaning. Therefore, Portuguese language used on QRH checklists can be considered difficult to read and understand word meaning, if participants are accustomed to using written English language. Lexical inferencing was also utilized to guess word meaning due to participants' inadequate background knowledge. This led to long response times, inadequate educated guesses to vocabulary word meanings, and inadequate responses on the flight deck to non-normal conditions (i.e. electrical and hydraulic faults). As participants' English language proficiency was adequate (high and medium levels), it is peculiar as to why they did not understand the meaning some abbreviations and acronyms in the notes section of the QRH checklist. Flight safety was also negatively impacted when participants utilized Portuguese language to solve electrical and hydraulic faults. It was indicated that long response times impacted their ability to recover the aircraft from electrical and hydraulic faults. Fault recovery technique was negatively impacted and thus other un-related to the fault, routine tasks (normal conditions) were abandoned due to difficulties with reading and understanding the Portuguese translated checklists. Design and integration of written English language vocabulary word types are predicated on the fact that participants must have background knowledge on these types of words. When written English language words were translated into Portuguese language, it negatively impacted interpretation of information in Portuguese language. As ESL flight crewmembers indicated they unilaterally translate written English language into their native language, it was obvious to the researcher to translate English language into their native language, therefore making it easier for ESL flight crewmembers to read and comprehend text on the ECAM and QRH checklists, in the researcher's experiment. Considering these factors, the researcher expected to find a significant positive correlation between participants NASA TLX workload scores and their response time when they read and comprehend technical information on the ECAM (written English language) Portuguese language QRH checklists. This outcome was likely due to participant's lack of background knowledge with QRH checklists translated into their native language, and due to their English language proficiency and metacognitive strategies utilized to read and comprehend information on the written Portuguese language QRH checklists. The researcher expected to find a positive correlation between participant NASA TLX workload scores and their response time when they read and comprehend technical information on ECAM (written English language) written English language QRH checklists. However, there was not a significant positive correlation between the two variables. Therefore, the data is not generalizable to the population of ESL flight crewmembers. As previously discussed, this outcome was likely due to participant's minimal difficulty they experienced while using written English language on the ECAM and QRH checklists. Their background knowledge, English language proficiency, and metacognitive strategies enabled them to perform well. Two-way ANOVA analysis revealed no significant main effect and interaction observed between ESL participant years of experience and English language proficiency and their reaction time, when the read and comprehend written English language on crew alerting systems and QRH checklists. This is an opposite finding from the researcher's expectations. However, there are a number of factors that help explain these results. First of all, participants had a range of airline experience levels and experience related to background knowledge reading and comprehending written English language on crew

alerting systems and QRH checklists. They were familiar with design and integration of written English language on crew alerting systems and QRH checklists. Participant familiarity with written English language design and integration on crew alerting systems and QRH checklists enabled them ability to understand text during the experimental trials. Second, there were participants that utilized metacognitive strategies to read and understand written English language. This may have helped them process information adequately during the experimental trials. Participant proficiency levels were adequate, and this could have also impacted their performance. As the researcher separated participant airline experience into two levels (20 years or greater versus less than 20 years), having less than 20 years of airline experience with high level of proficiency resulted in faster response times to crew alerts. On the other hand, there were some participants that had a long response time to crew alerts with medium level proficiency. Participants with 20 years of experience and greater with high level of English language proficiency responded slower to crew alerts than medium level participants. Participant number of airline years of experience does not appear to be a factor with a significant main effect on participant reaction time. Perhaps, background knowledge and training may be more efficient variables to research without specific numerical value focus (i.e. less than 20 years of airline experience, 20 years or greater of airline experience) in future research. As this experiment measured flight crewmember performance that were Portuguese natives, it would seem practical to test other flight crewmembers that have an array of linguistic backgrounds. Results could be different if testing participants with other linguistic backgrounds (e.g. Mandarin) during experimental trials, and may convey an interaction between aforesaid variables. Literature review indicated high/medium proficiency level participants use different strategies to read and comprehend written English language. There were participants that indicated they were highly proficient with reading and comprehending written English language, and aware of strategies to use while reading and comprehending written English language. They also indicated they were challenged with terminology on crew alerting system and QRH checklists. As Yildiz-Genc (2009) indicated, background knowledge and English language proficiency is a factor that influences ESL adults' ability to read and comprehend written English language. In the researcher's experimental study, participant proficiency levels were high and medium and they had adequate background knowledge in the text they read and comprehended during the trials. Therefore, this finding corroborates Yildiz-Genc (2009) finding that differences with participant English language proficiency are expected when they read and comprehend written English information. If the researcher had imposed a time limitation on the trials, the results may have been different. As Hashemi and Bagheri's (2014) study indicated, no time limit resulted in better comprehension of texts, whereas a time limit had a negative impact on performance. The researcher's finding corroborates Hashemi and Bagheri's (2014) study. Second two-way ANOVA also indicated no significant main effect and interaction between participant English language proficiency and NASA TLX workload scores, when they read and comprehend information on crew alerting systems and QRH checklists. Crew alerting systems and QRH checklists that were analyzed contained text genre that was technical/scientific and text corpora contained high number of high frequency words and academic words, this likely had an positive effect on flight crewmember ability to read and understand text

on crew alerting systems and QRH checklists. Coxhead (1998) and West (1953) indicated that high frequency words and academic words in text have a higher comprehensibility than other words (e.g. low frequency). Participants in the researcher's study had background knowledge, years of experience, and training with technical words on crew alerting systems and QRH checklists. This likely reduced participant cognitive workload, enabled them to recognize, read and comprehend technical words, while perform tasks during non-normal conditions. Wanpen et al. (2013) study indicated that participant technical vocabulary knowledge helped participants with reading text. As Mehrpour and Riazi (2004) indicated, high proficiency, background knowledge in text is important when reading and comprehending different words in text corpora. As the researcher did not alter sentence length or simplify text (text was authentic), this could be the reason why participants performed well reading and comprehending written English language text on crew alerting systems and QRH checklists. On the other hand, there were participants that experience higher cognitive workload compared to other participants. This could be due to participants with high proficiency using metacognitive strategies.

6 Conclusion

Written English language on the ECAM system and associated QRH checklists did not have a substantial negative impact on ESL flight crewmembers' performance. But, other languages should be investigated to determine if this is an expectation of other regions, and flight crewmembers with different linguistic backgrounds across the globe. In other words, is the issue of written English language still a factor in other regions of the globe? Since the researcher's experiment focused on one region, other regions should be investigated as well. On the other hand, since translating English language into flight crewmembers' native language was an issue that impacted their performance, other regions and languages of flight crewmembers should be included in future research studies.

Acknowledgments. The author is a Sr. Scientist and Engineer for the Boeing Company. "The views expressed in this article are solely those of the author in a private capacity and do not in any way represent the views of The Boeing Company".

References

- Abdul-Hamid, S., Samuel, M.: Reading scientific texts: some challenges faced by EFL readers. *Int. J. Soc. Sci. Hum.* 2(6), 509 (2012)
- Al-Sohbani, Y., Muthanna, A.: Challenges of Arabic-English translation: the need for re-systematic curriculum and methodology reforms in Yemen. *Acad. Res. Int.* 4(4), 442 (2013)
- Baker, L., Brown, A.L.: Cognitive monitoring in reading. In: Flood, J. (ed.) *Understanding Reading Comprehension*, pp. 21–44. International Reading Association, Newark (1984)

- Barani, M., Karimnia, A.: An investigation into translation students' english reading comprehension skills and strategies: a cross-sectional study. *Elixir Linguit. Trans.* **73**, 26257–26262 (2014)
- Bureau d'Enquetes et d'Analyses (BEA): Final Report on the accident 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight 447 Rio de Janeiro-Paris (2012)
- Center for Investigation and Accident Prevention (CENIPA): CENIPA Final Report on Accident involving L410 Noar Linhas Aereas Flight 4896 13 July 2011; RF A-019/CENIPA/2013 (2013)
- Coady, J., Huckin, T.: *Second Language Vocabulary Acquisition: A Rationale For Pedagogy*. Cambridge University Press, Cambridge (1997)
- Chung, T.M., Nation, P.: Identifying technical vocabulary. *System* **32**(2), 251–263 (2004)
- Condelli, L., Wrigley, H.S.: Instruction, language and literacy: What works study for adult ESL literacy students. *LOT Occas. Ser.* **6**, 111–133 (2006)
- Coxhead, A.J.: *An Academic Word List*. English Language Institute Occasional Publication Number 18. Victoria University of Wellington, Wellington (1998)
- Drury, C.G., Ma, J.: *Language Errors in Aviation Maintenance: Final Report*. Reports to William J. Hughes Technical Center, the Federal Aviation Administration under research grant #2002-G-025 (2005)
- Dycus, D.: Guessing word meaning from context: should we encourage it? *Lit. Across Cult.* **1**(2), 1–6 (1997)
- Dyson, M.C.: How physical layout affects reading from screen. *Behav. Inf. Technol.* **23**(6), 377–393 (2004)
- Dwaik, R.A., Shehadeh, A.M.: Guessing patterns of palestinian college students. *Read. Matrix* **13**(1), 14–26 (2013)
- FAA Human Factors Team: *The interfaces between flight crews and modern flight deck systems*. Federal Aviation Administration, Washington, D.C. (1996)
- FAA: *Operational use of flight path management systems*. Report of the Performance based operations Aviation Rulemaking Committee/Commercial Aviation Safety Team Flight Deck Automation Working Group (2013)
- Hashemi, S.Z., Bagheri, M.S.: External factors influences on EFL learners reading comprehension test performance. *Int. J. Lang. Learn. Appl. Linguist. World (IJLLALW)* **7**(1), 150–166 (2014)
- Hancock, O.H.: *Reading Skills for College Students*, 4th edn. Prentice Hall, Upper Saddle Rivers (1998)
- Hartley, J.: Designing instructional text for older readers: a literature review. *Br. J. Educ. Technol.* **25**(3), 172–188 (1994). National Council for Educational Technology
- Ho, L.-C.: *A critical analysis of airline safety management with reference to pilots and aviation authority officers*. Ph.D. Dissertation, Cranfield University, December 1996
- Hutchins, E., Nomura, S., Holder, B.: The ecology of language practices in worldwide airline flight deck operations: the case of Japanese airlines. In: *Proceedings of International Conference on Human-Computer Interaction in Aeronautics*, Seattle, WA, September 2006, pp. 290–296 (2006)
- Interstate Aviation Committee Air Accident Investigation Commission (IAC): *Final Report on Boeing 737-505 VP-BKO Aircraft Accident* (2009)
- Karbalaei, A.: A comparison of the metacognitive reading strategies used by EFL and ESL readers. *Read. Matrix* **10**(2), 165–180 (2010)
- Kim, H.: Effects of Korean students' contextual and lexical knowledge on L2 text comprehension. *Engl. Teach.* **61**(3), 83–103 (2006)

- Kobayashi, H., Rinnert, C.: Effects of first language on second language writing: translation versus direct composition. *Lang. Learn.* **42**, 183–215 (1992)
- Komite Nasional Keselamatan Transportasi (KNKT): Aircraft Accident Investigation Report: PT. Indonesia Air Asia Airbus A320-216; PK-AXC; Karimata Strait. Republic of Indonesia; 28 December 2014 (2015)
- Mali Commission of Inquiry (MCI): Ministère De L'Équipement, Des Transports ET DU Désenclavement. Commission D' Enquete Sur Les Accidents Et Incidents D'Aviation Civile. Final Report. Accident on 24 July 2014 MD-83 registered EC-LTV operated by Swiftair S.A. (2016)
- Mehrpour, S., Riazi, A.: The impact of text length on reading comprehension in English as a second language. *Asian EFL J.* **3**(6), 1–14 (2004)
- Nation, I.S.P.: *Learning Vocabulary in Another Language*. Cambridge University Press, Cambridge (2001)
- Nylander, E.: *The inferencing behaviour of Swedish EFL university students: a quantitative analysis of lexical inferencing in relation to vocabulary depth*. Lund University: Centre for Languages and Literature (2014)
- Ogilvie, G.: *The impact of culture on communications: a study on the possible effects of culture on inter-cockpit communications*. University of Hong Kong (1984)
- Park, J., Yang, J.S., Hsieh, Y.C.: University level second language readers' online reading and comprehension strategies. *Announcements & Call for Papers*, p. 148 (2014)
- Parry, K.: Building a vocabulary through academic reading. *Tesol Q.* **25**(4), 629–653 (1991)
- Salle, G.P., Gibbons, D.M.: AIA/AECMA Project Report on Propulsion System Malfunction Plus Inappropriate Crew Response (PSM + ICR), vol. 1, 1 November 1998 (1998)
- Sevillian, D.B.: *Flight deck engineering: impact of flight crew alerting and information systems on English as a second language flight crewmembers performance in airline flight operations*. Cranfield University 2017 CERES Published Dissertation, UK (2017)
- Smith-Jackson, T.L., Wogalter, M.S.: Applying cultural ergonomics/human factors to safety information research. In: *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society*, vol. 6, pp. 150–154. Human Factors Society, Santa Monica (2000)
- Wanpen, S., Sonkoontod, K., Nonkukhetkhong, K.: Technical vocabulary proficiencies and vocabulary learning strategies of engineering students. *Procedia-Soc. Behav. Sci.* **88**, 312–320 (2013)
- Wang, A.G.: A methodological probe to aeronautical english vocabulary instruction. *Open J. Mod. Linguist.* **1**(2), 45–51 (2011a)
- Wang, Q.: Lexical inferencing strategies for dealing with unknown words in reading - a contrastive study between Filipino Graduate students and chinese graduate students. *J. Lang. Teach. Res.* **2**, 302–313 (2011b)
- West, M.: *A General Service List of English Words*. Longman, Green & Co., London (1953)
- Wogalter, M.S., Begley, P.B., Scancorelli, L.F., Brelsford, J.W.: Effectiveness of elevator service signs: measurement of perceived understandability, willingness to comply and behavior. *Appl. Ergon.* **28**, 181–187 (1997)
- Yildiz-Genc, Z.: An investigation on strategies of reading in first and second languages. *Selected papers from 18th ISTAL*, pp. 407–415 (2009)
- Ynfiesta, B., Suarez, L., Fernandez-Peraza, A.: Translation of acronyms and initialisms in medical texts on cardiology. *Cuba. Soc. Cardiol: CorSalud* **2013** **5**(1), 93–100 (2013)
- You, C.: A comparative study of second language reading comprehension from paper and computer screen (2009)
- Zhao: Using translation in ESL classrooms: an Asian perspective. *Int. J. Innov. Interdiscip. Res.* Chengdu Institute Sichuan International Studies University, China (2015). V2 14 2015



The Preliminary Application of Observer XT (12.0) in a Pilot-Behavior Study

Ruishan Sun^(✉), Guanchao Zhang, and Zhibo Yuan

Research Institute of Civil Aviation Safety, Civil Aviation University of China,
Tianjin, China

sunrsh@hotmail.com, guanchaozhang@outlook.com,
zhibiyuan@163.com

Abstract. In order to study pilots' behavior characteristics, two pilots of a certain airline were selected as research subjects. Two typical tasks in recurrent training were selected for the experimental scene. One was an aerodrome-traffic pattern under a normal situation; the other was an aerodrome-traffic pattern in the case of a large crosswind. Using multi-angle video recording, all details of the two pilots' operation in the training simulator (B737-800) were recorded completely. Using Noldus's Observer XT 12.0, a preliminary analysis of typical operational behaviors was performed, including the control behaviors of the pitch, yaw, and roll movement, as well as the throttle lever movement. The coding scheme and the data visualization of these behaviors were also presented. Finally, combing the statistics, a depth-comparison analysis of these behavior characteristics was conducted in terms of many aspects, including mean duration, total number, rate per minute, percentage of total duration, and so on. The results show that the pilot's pitch and roll controls have larger differences in mean duration, total number, rate per minute, and percentage of total duration; however, there were no significant differences in other behaviors between tasks.

Keywords: Pilots' behavior characteristics · The Observer XT

1 Introduction

“Aviation safety” is a topic of perpetual research in the aeronautical field. Government administrations, aircraft manufacturers, and airlines have been working hard to improve the safety of aircraft. Whether it is structural improvements, new electronic devices, or new means of communication, the aim is to pursue higher security and to maximize economic interests in ensuring safety [18]. The reliability of aircraft has been greatly improved, which profited from the development of aviation design and manufacturing and aviation safety has increasingly considered human factors. Statistics from different sources indicate that crew errors have always been the main cause of civil-aviation accidents [1]. Therefore, it is of great significance to study the operational behaviors of pilots and to analyze their characteristics during flight.

There have been many pilot-behavior studies. Internationally, Bonomalenko et al. [2] considered operational behavior as elements of pilot action in his book, *Flight Psychology*, and summarized pilot operation in terms of integrity, accuracy, timeliness,

flight image, and a series of features. Liu and Liu [12], using a developed psychological scale for civil-aviation pilots, designed a flight-behavior-observation system using VB and Access to realize the functions of testing, result inquiry, data management, statistical analysis, and user management and provide an experimental platform to psychologically assess pilots and pilot cadets. Hayashi et al. [8] built up a neural-network model based on genetic-algorithm optimization using a simulator to obtain flight data, and analyzed the pilot's behavior in terms of the sensitivity and threshold of this model. The results revealed the operational-behavior rule of the pilot. Chen and Wang [3] proposed a unique frequency-domain-analysis method based on simulated experimental data to reflect the pilot's activity frequency and activity level after setting the cut-off and power frequencies of pilot flight-control behavior as key indicators. Smith et al. [19] introduced common methods and techniques for pilot-behavior modeling. Keane [11] proposed an extended Lancheste equation-evaluation method based on partial differential equations. Hillard et al. [9] proposed a pilot-behavior-assessment method, mainly used in the field of information extraction.

Chinese scholars have also been involved in this research. He et al. [7] analyzed flight accidents and incidents caused by crew error between 1996 and 2000 based on records from the Civil Aviation Administration of China and selected cases that are closely related to the time margin for problem solving; they discussed these cases with flight experts, and drew a relationship between such incidents and crew behavior and time margin. Yin et al. [23] proposed an air-combat-pilot fighting-behavior-assessment method based on average time of air combat, dominant posture, and air-combat credibility. Chen and Tan [4] used the principle of EMG (Electromyography) and STP (Skin Temperature) detection to design pilot-behavior-analysis experiments based on electromyography and skin-temperature testing using JD/PW-5 testing equipment and the PC-based aviation-training device. Xue et al. [22] used the ACT-R (Adaptive Control of Thought-Rational) cognitive framework to model the internal mechanism for obtaining, extracting and applying the skills of civil-aircraft pilots and structured simulation modeling of behavioral integration. Luo et al. [14] discussed the relationship between psychological factors such as social-psychological quality, motivation, emotion, and personality psychology and crew-behavior errors and analyzed the psychological background of such errors; the relationship among the flight-space environment, aviation-organization management, the influence of man-machine-environment imbalance, and the influence of crew mismanagement was also discussed, and countermeasures to improve crew management were also proposed. Wu and Wang [21] proposed serial process hypothesis of human brain, set RNP APCH profile as operation scenario background, translates flight crew operation behavior into abstract mathematic model and quantitatively produces the level of dependence and strength of workload utilizing mathematic means. The work intensity of the operation and the correlation between operational tasks were quantitatively given by mathematical methods, which better described the real human-computer interaction. Liu [15] studied the operational-gesture characteristics of pilots' intelligent model using visual-monitoring technology. The contents covered were machine-vision-based gesture detection, tracking, trajectory analysis and cockpit operational behavior analysis combined with the eye-movement characteristics of the pilots. At present, the domestic and foreign scholars' research on pilot behavior mainly focuses on pilot behavior psychology

research, pilot behavior modeling research and pilots behavior assessment. The accuracy and applicability of the research conclusions need to be improved. However, there are few researches on the basic operational behaviors between pilots in different flight training subjects through the direct behavior observation using dynamic flight simulator.

This paper uses wireless cameras and monitoring equipment in the B737-800W Full Flight Simulator to construct an experimental platform for pilot-behavior observation. Then, in training tasks under typical aerodrome-traffic patterns in the normal situation and with large crosswinds, the pilots' behaviors are videotaped. Using the observation and analysis function of Observer XT (12), this paper aims to explore the similarities and differences in pilot behaviors under the two kinds of training tasks.

2 Method

2.1 Participants

The study involved two male-refreshment pilots (33-year-old captain, 4554-h flight experience; 26-year-old copilot, 900-h flight experience) at an airline who had good flying skills, normal vision, and good physical condition. As the data acquisition involved the human subject, this experiment was approved by the Ethics Review Committee of Civil Aviation University of China. Two pilots read the informed consent form and voluntarily signed and then participated in the trial before starting the experiment.

2.2 Apparatus

- 2.2.1 An airline's B737-800 W Full Flight Simulator, as shown in Fig. 1-a, is mainly used for pilots' regular refreshment. It can simulate a variety of flight missions realistically, letting the pilots act as if they were manipulating controls on a real plane;
- 2.2.2 EZVIZ surveillance video equipment and four EZVIZ wireless cameras, as shown in Figs. 1-b and c, are used to record pilots' manipulative behaviors in the cockpit;
- 2.2.3 Four camera bases made in-house and three tripods, as shown in Figs. 1-d and e, are mainly used to fix the camera flexibly.
According to the process shown in Fig. 1, the existing equipment is connected together to build an experimental platform suitable for observing the pilot's behavior, as shown in Fig. 1-f.
- 2.2.4 Observer XT (12.0) - Behavioral Analysis Software: In order to be able to quantitatively analyze pilots' operational behavior, Noldus' Observer XT 12.0 Behavioral Analysis Software is used. Unlike conventional behavioral observation devices, it can be used to record and analyze the actions of the studied subjects, Attitude, emotion, social interaction, human-computer interaction and so on. It is a standard tool for studying human behaviors to record the times, occurrences and durations of various behaviors of the subjects under study.

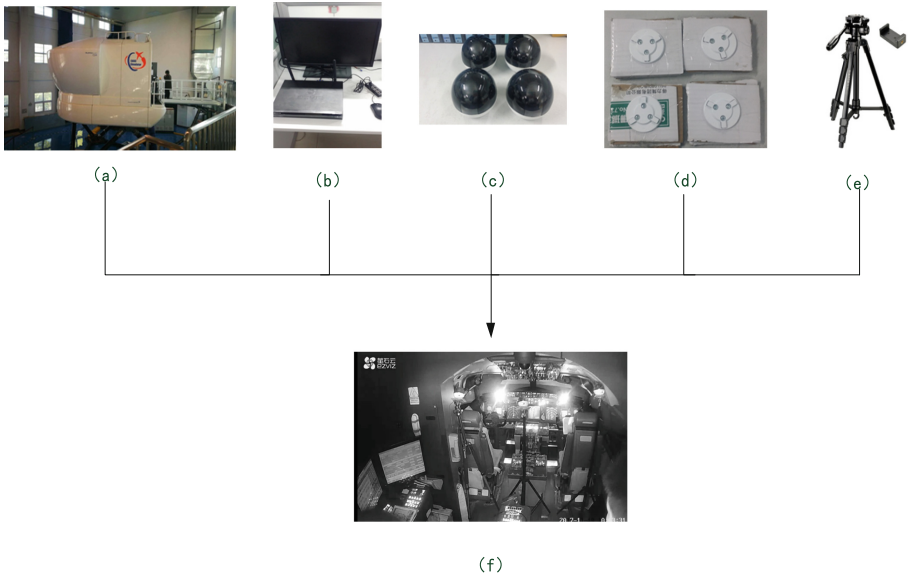


Fig. 1. Components of the experimental platform

2.3 Research Design

Based on an airline’s B737-800W Full Flight Simulator, an experimental platform was constructed to observe pilot operations using existing equipment, and experimental design was carried out. Afterwards, the refreshment-manipulation videos of the pilots were obtained from the experimental platform in the Full Flight Simulator.

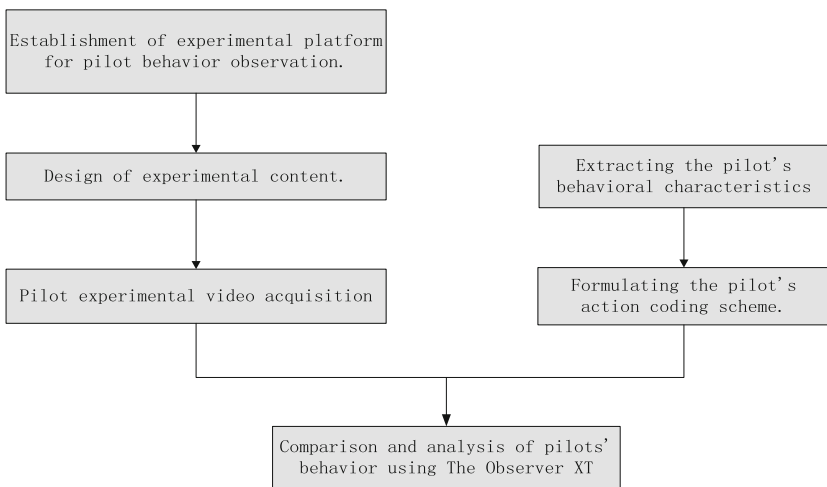


Fig. 2. Study framework

Before importing the maneuvering video into the Observer XT, the pilot's motion-coding scheme needed to be pre-defined. Then, the pilots' behaviors were analyzed by the developed coding scheme and compared under different tasks. Specific research ideas are shown in Fig. 2.

2.4 Data Collection and Analysis Processes

2.4.1 Data Collection Processes

- (1) The B737-800 W Full Flight Simulator's Cockpit and the surveillance-video equipment are activated.
- (2) The matters needing attention in this experiment are explained.
- (3) An aerodrome-traffic circuit under the normal situation is set up on the control computer by the pilot instructor, and the airplane is set directly on the runway end, eliminating taxiing from the air bridge to the end of the runway and simplifying the experimental process.
- (4) After the pilot instructor has issued the "take-off" command, the two pilots operate according to the established procedure and route, which is the only time that communication between the crew members and between the flight crew and the controller is allowed; silence should be maintained for the rest of the process so as not to affect the pilots' normal operation.
- (5) The task of the aerodrome-traffic pattern ends under the normal situation, when the plane lands on the runway and comes to a stop.
- (6) After three minutes of rest, the pilot instructor sets up an aerodrome-traffic pattern of large crosswind on the control computer and pulls the plane directly to end of the runway.
- (7) After the pilot instructor has issued the "take-off" command, the two pilots should operate according to the established procedure and route, which is the only time that communication between the crew members and between the flight crew and the controller is allowed; silence should be maintained for the rest of the process so as not to affect the pilots' normal operation.
- (8) The aerodrome-traffic-pattern task ends under large crosswind when the plane lands on the runway and comes to a stop.
- (9) The surveillance-video equipment is deactivated, and then it is removed and packaged.

Matters needing attention: before starting the experiment, it is important to ensure that all experimental instruments are working properly and that the camera can capture clear video of the pilots' operations.

2.4.2 Analysis Processes

- (1) Coding and defining the pilot's behavior

Before observing the record, the pilot's behaviors need to be coded. By reading the "*Airplane Flying Handbook*" [6] and watching the "*Pilots Eye*" videos, pilots' characteristics, including Holding, Landing-gear Setting, Throttle Control, Steering-column

Control, Flap Setting, Speed-brake Setting, Rudder Setting, and Thrust-reverser Setting, are all extracted.

Based on the pilots’ extracted behavioral characteristics and combined with the code rules of the ‘Codings’ of the Observer XT software instructions, the pilot’s behaviors are divided into four continuous behavior groups (including Pitch Control, Roll Control, Yaw Control, and Monitoring) and a start–stop behavior group (Other behaviors). Behavior groups and specific definitions of behaviors are shown in Table 1.

Table 1. The pilot’s coding scheme

| The type of the behavior group | The name of the behavior group | Behavior | Detailed description |
|--------------------------------|--------------------------------|------------------------------|---|
| Continuous behavior groups | Pitch control | Pull back on the stick | In the pitch control direction, the joystick is shifted from the static State after pushing forward or neutral static state into the back pull state until the stick-forward movement just to takes place |
| | | Push the stick forward | In the pitch control direction, the joystick is shifted from the static state after pulling backward or neutral static state into the forward push state until the stick-back movement just to take place |
| | | Keep pitch neutral | The joystick is in neutral position on the pitch control direction |
| | Roll control | Compressive bar to the left | In the rolling control direction, the joystick is shifted from the static state after compressing bar to the right or neutral static state into the compressing bar to the left state until the compressing bar to the right occurs |
| | | Compressive bar to the right | In the rolling control direction, the joystick is shifted from the static state after compressing bar to the left or neutral static state into the compressing bar to the right state until the compressing bar to the left occurs |
| | | Keep the roll neutral | The joystick is in neutral position on the roll control direction |
| | Yaw control | Left rudder pedal | In the yaw control direction, the rudder is shifted from the static state after the right rudder pedal or neutral static state into the left rudder pedal state until the right rudder pedal action occurs |

(continued)

Table 1. (continued)

| The type of the behavior group | The name of the behavior group | Behavior | Detailed description |
|--------------------------------|--------------------------------|--|---|
| | | Right rudder pedal | In the yaw control direction, the rudder is shifted from the static state after the left rudder pedal or neutral static state into the right rudder pedal state until the left rudder pedal action occurs |
| | | Keep heading neutral | The joystick is in the neutral position in the yaw control |
| Start-stop behavior group | Other behaviors | Push the throttle lever forward | Turn the throttle lever from static condition to pushing forward condition until it keeps the stationary state again |
| | | Pull the throttle lever back | Turn the throttle lever from static condition to pulling back condition until it keeps the stationary state again |
| | | Turn on the reverse thrust | A process that the PF turns the throttle lever from static to back until it stops again after turning on the reverse thrust switch |
| | | Set the flaps | The PF turns the flap lever from one static position to another static position |
| | | Retract the landing gear | A process that the PNF turns the landing gear from the droop position to the retracted position |
| | | Place the landing gear in OFF position | PNF puts the landing gear handle OFF position |
| | | Lower landing gear | A process that the PNF turns the landing gear from the retracted position or the OFF position to the droop position |
| | | Speedbrake Arming | Place the speedbrake in the position of Arming |
| | | Put down the speedbrake | A process of returning the speedbrake to the original position |
| | | Release the parking brake | A process of turning the parking brake from the ON position to the OFF position |
| | | Open the parking brake | A process of turning the parking brake from the OFF position to the ON position |

(2) Behavioral observation records

Custom behavior codes are applied to observe and record the pilots’ manipulations under the two tasks and Fig. 3 presents a screenshot. In Fig. 3, the observation time of the aerodrome-traffic pattern under the normal situation is 792.95 s, and that under large crosswind is 456.16 s.

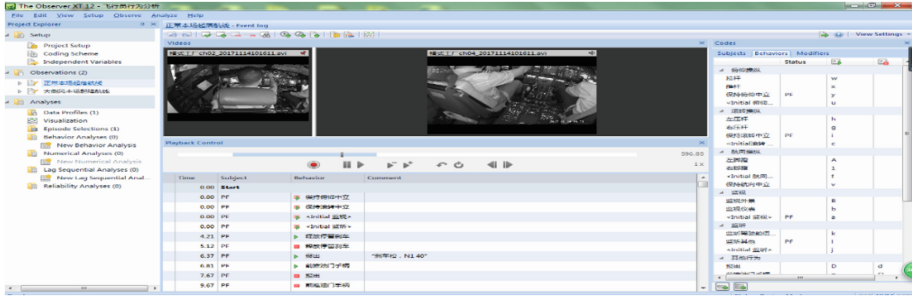


Fig. 3. Screenshot of the pilot’s behavior observation record

(3) The mean duration, total number and the proportion of each behavior under the two tasks

By clicking the “Calculate” option in the Behavior Analysis Function, one can count the total durations of each pilot behavior, as well as the total number of occurrence. Because the time consumption of the two flight tasks is inconsistent, it is not appropriate to directly compare the total duration and total number of each operational behavior. Therefore, we need to transform the data into the same standard form using formulas (1), (2) and (3), and then compare them [5].

Formula (1) is used to calculate the number of occurrences of each behavior per minute, known as the rate per minute (RPM):

$$RPM = \frac{\text{the total number of some kind of behavior}}{\text{Observation time}} * 60. \quad (1)$$

Formula (2) is used to calculate the total duration of each behavior as a percentage of the total observed duration:

$$\text{The proportion of each behavior} = \frac{\text{The total duration of each behavior}}{\text{Observation time}} * 100\%. \quad (2)$$

Formula (3) is used to calculate the mean duration of each behavior:

$$\text{The mean duration of each behavior} = \frac{\text{The total duration of each behavior}}{\text{The total number of each behavior}}. \quad (3)$$

3 Results and Discussions

Pilot control behaviors are directly affected by the internal and external environment and the controlled object [4]. This study is different from previous studies which adopt expert investigation and literature review to design flight control behavior training and evaluation tools [14, 23], use semi-quantitative method for flight crew resource management behavior and workload evaluation [13], use computer technology to the pilot manipulation simulation experiment to research (see, for example, [22]) and so on. It uses the Observer XT (12) to quantitatively describe the process of pilots' controlling behaviors of aerodrome traffic patterns in the normal situation and the large crosswind, in ensuring data acquisition under that the condition of internal and external environment is not affected to find out the specific differences between the pilots' operating behaviors of the two tasks in the cockpit. The following results and discussions are made from the differences in visibility charts of the behavioral data, Rolling control, pitching control, and other behaviors.

3.1 Visibility Charts of the Behavioral Data

After observing and recording the pilots' manipulation video, the "Visual" option of the software can automatically generate the visibility charts of aerodrome-traffic patterns in the normal situation and in that with large crosswinds, as shown in Figs. 4 and 5. In the generated visibility charts, each rectangular strip represents the occurrence of an encoding action, and the length of each rectangle represents its duration [25].

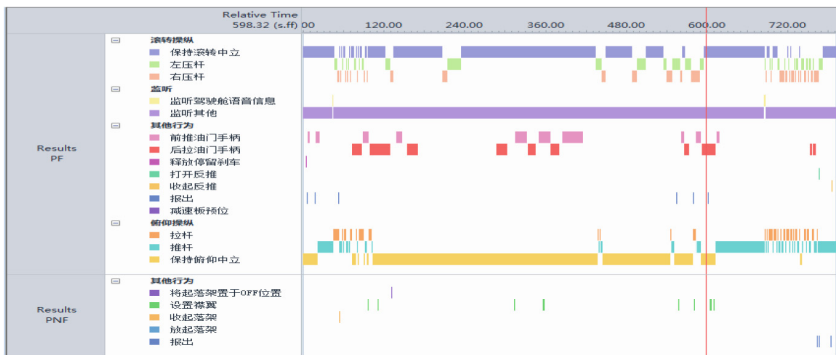


Fig. 4. Visibility Chart of the pilots' behaviors in the normal aerodrome-traffic pattern

Under a normal aerodrome-traffic pattern, it can be seen from the pilot-behavior-visibility chart that in the two time periods of 0–180 ms and 660–790 ms, there are many more rectangular strips in the roll-control and pitch-control behavior groups than between these periods. This also means that pilots manipulate the pitch and roll of the aircraft during the takeoff and landing phases more frequently.



Fig. 5. Visibility Chart of the pilots’ behaviors in the large-crosswind aerodrome-traffic pattern

Under the large-crosswinds aerodrome-traffic pattern, the chart shows that the pilot exerts more control over the pitch and roll of the aircraft than in the normal case over the whole period.

Through the data analysis process, the statistic tables of each behavior data of pilots under two different tasks are generated, as shown in Tables 2 and 3, respectively.

Table 2. Statistics of pilot behaviors under normal aerodrome traffic

| | Behavior | Total number (time) | RPM | Total duration (s) | Mean duration (s) | Proportion |
|-----|--|---------------------|------|--------------------|-------------------|------------|
| PF | Keep the roll neutral | 27 | 2.04 | 559.24 | 20.71 | 70.53% |
| PF | Compressive bar to the left | 47 | 3.56 | 134.99 | 2.87 | 17.02% |
| PF | Compressive bar to the right | 45 | 3.41 | 98.73 | 2.19 | 12.45% |
| PF | Push the throttle lever forward | 10 | 0.76 | 108.86 | 10.89 | 13.73% |
| PF | Pull the throttle lever back | 11 | 0.83 | 136.10 | 12.37 | 17.16% |
| PF | Release the parking brake | 1 | 0.08 | 0.92 | 0.92 | 0.12% |
| PF | Pull back on the stick | 53 | 4.01 | 83.46 | 1.57 | 10.53% |
| PF | Push the stick forward | 52 | 3.93 | 187.42 | 3.60 | 23.64% |
| PF | Keep pitch neutral | 13 | 0.98 | 522.06 | 40.16 | 65.84% |
| PF | Turn on the reverse thrust | 1 | 0.08 | 1.25 | 1.25 | 0.16% |
| PF | Retract the thrust reverser | 1 | 0.08 | 1.94 | 1.94 | 0.24% |
| PF | Speedbrake Arming | 1 | 0.08 | 1.08 | 1.08 | 0.14% |
| PNF | Set the flaps | 8 | 0.61 | 12.84 | 1.61 | 1.62% |
| PNF | Retract the landing gear | 1 | 0.08 | 0.59 | 0.59 | 0.07% |
| PNF | Lower landing gear | 2 | 0.15 | 1.13 | 0.56 | 0.14% |
| PNF | Place the landing gear in 'OFF' position | 1 | 0.08 | 0.73 | 0.73 | 0.09% |

Table 3. Statistics of pilot behaviors under large-crosswind aerodrome traffic

| Object | Behavior | Total number (time) | RPM | Total duration (s) | Mean duration (s) | Proportion |
|--------|---|---------------------|-------|--------------------|-------------------|------------|
| PF | Keep the roll neutral | 22 | 2.69 | 82.57 | 3.75 | 18.10% |
| PF | Compressive bar to the left | 108 | 14.21 | 200.28 | 1.85 | 43.91% |
| PF | Compressive bar to the right | 111 | 14.60 | 173.31 | 1.56 | 37.99% |
| PF | Push the throttle lever forward | 7 | 0.92 | 17.42 | 2.49 | 3.82% |
| PF | Pull the throttle lever back | 14 | 1.84 | 23.20 | 1.66 | 5.09% |
| PF | Release the parking brake | 1 | 0.13 | 0.71 | 0.71 | 0.16% |
| PF | Pull back on the stick | 107 | 14.07 | 177.68 | 1.66 | 38.95% |
| PF | Push the stick forward | 112 | 14.73 | 197.77 | 1.77 | 43.36% |
| PF | Keep pitch neutral | 15 | 1.97 | 80.71 | 4.81 | 17.69% |
| PF | Turn on the reverse thrust | 1 | 0.13 | 0.90 | 0.90 | 0.20% |
| PF | Retract the thrust reverser | 1 | 0.13 | 3.62 | 3.62 | 0.79% |
| PF | Speedbrake Arming | 1 | 0.13 | 0.49 | 0.49 | 0.11% |
| PNF | Set the flaps | 2 | 0.26 | 2.46 | 1.23 | 0.54% |
| PNF | Retract the landing gear | 1 | 0.13 | 0.88 | 0.88 | 0.19% |
| PNF | Lower landing gear | 1 | 0.13 | 0.85 | 0.85 | 0.19% |
| PNF | Place the landing gear in OFF' position | 0 | 0 | 0 | 0 | 0 |

3.2 Rolling Control Differences

The roll control group includes the compressive bar to the left, the compressive bar to the right, and roll neutral manipulation. According to Fig. 6, it is found that the proportion of keeping neutral roll in the case of large crosswinds is 18.10%, compared with 70.53% under normal situations, which is obviously less than that under normal situations. In the behavior of the compressive bar to the left and right, the compressive bar to the left and right accounted for 43.91% and 37.99% respectively of the large crosswind condition. However, under normal situations, the compressive bar to the left and right accounted for 17.02% and 12.45%. It can be seen that in the case of large crosswind, the behavior of the compressive bar to the left or right is significantly more than normal. By comparing the RPMs of the behaviors in the roll-control group in Tables 2 and 3, it can be seen that the RPMs of behaviors to maintain roll-neutral behavior do not differ significantly between the two tasks, respectively 2.04 times per minute in normal situation and 2.69 times per minute in the large crosswind, whereas the RPM of moving the compressive bar to the left or right is obviously higher in the case of a large crosswind. Combining the above two points, it can be found that in the case of a large crosswind, both the proportions of behaviors' duration and the RPM

devoted to moving the compressive bar to the left and right are larger than in the normal situation.

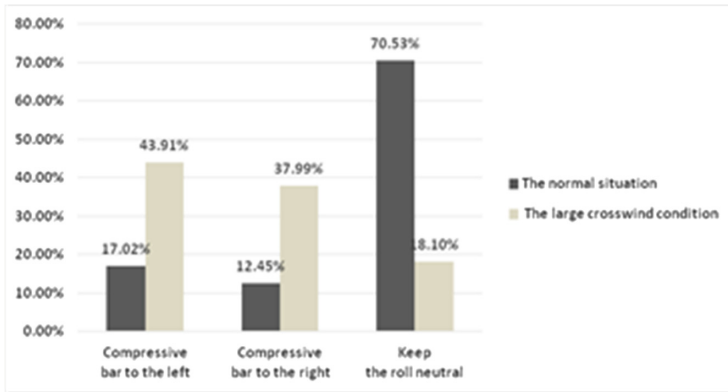


Fig. 6. The comparison of pilots’ roll-control behaviors for each of the two tasks

3.3 Pitching Control Differences

The pitch control group includes the pulling back on the stick, the pushing the stick forward, and pitch neutral manipulation. According to Fig. 7, it is found that the proportion of keeping neutral pitch in the case of large crosswinds is 14.77%, compared with 65.84% under normal situations, which is obviously less than that under normal situations. In the behaviors of the pulling back on the stick and the pushing the stick forward, the pulling back on the stick and the pushing the stick forward accounted for 38.95% and 43.36% respectively of the large crosswind condition. However, under normal situations, the pulling back on the stick and the pushing the stick forward

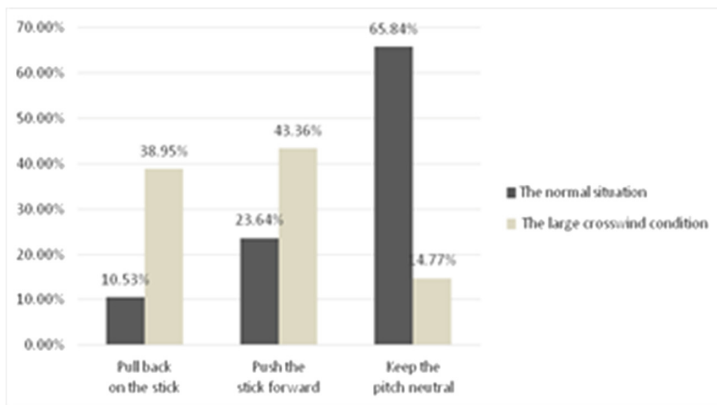


Fig. 7. The comparison of pilots’ pitch-control behaviors for each of the two tasks

accounted for 17.02% and 12.45%. It can be seen that in the case of large crosswind, the behavior of pulling or pushing the stick is significantly more than normal. By comparing the RPMs of each behavior in the pitch-control groups in Tables 2 and 3, one finds that RPMs that pitch-neutral-maintaining behaviors do not differ significantly between the two tasks, whereas pushing and pulling of the throttle lever occur with greater frequency in the large-crosswind case. Combining the above two points, it can be found that in the case of large crosswind, the throttle lever is pushed and pulled more frequently and for higher proportions of behaviors' duration.

3.4 Other Behaviors

Through the calculation, the statistics for each pilot behavior under the two tasks are generated, as shown in Tables 2 and 3. Comparison of the tables shows that, in other behavior groups, the proportions of six behaviors (Release parking brake, Close the landing gear, Landing Gear Down, Speed-braker Arming, Turn on the Reverse Thrust and Retract the Reverse Thrust) are less than 1% for both tasks and there is no difference in the frequency of occurrence.

The process of aircraft manipulation is a closed-loop human-machine interaction process including the display and acquisition process of task information, the cognition process of pilots, the execution process of mission actions, and the information processing process of flight control computer. Aiming at the process of aircraft manipulation, Liu used the basic elements of colored Petri nets to describe the manipulative dynamic process, built the manipulation process model based on colored Petri nets and put forward the corresponding task reachability, cognitive load and entropy-based method of evaluating the complexity of the program [17]. Unlike above, this paper selected the task execution process and referred to that Sun et al. proposed the idea of quantitatively evaluating the pilot's operation level [20]. The pilot's operation behaviors in the cockpit are divided into continuous behavior groups (Pitch Control, Roll Control, Yaw Control) and start-stop behavior groups (Release parking brake, Close the landing gear, Landing Gear Down, Speed-braker Arming, Turn on the Reverse Thrust, Retract the Reverse Thrust and so on). And then each operational behavior is defined, and a coding scheme is constructed. With the help of Observer XT (12.0) and statistical methods, the collected videos are quantitatively recorded to analyze the total number, mean duration, total duration, the number of occurrences per minute and the proportion of each behavior duration of pilots during two tasks. Compared with a previous study on pilots' verbal behavior by using Observer XT software, it enriches the behavior code of pilots, records pilots' manipulation behavior from multiple angles, and conducts in-depth research on pilots' main operation behaviors. Through the above the results show that compared with the normal aerodrome-traffic pattern, the RPMs and proportions of four kinds of behavior (Pull back on the stick, Push the Stick forward, Compressive bar to the left, Compressive bar to the right) are significantly more and operations are more complicated under large-crosswind aerodrome traffic; the total number of occurrence and proportions of six kinds of behavior (Release parking brake, Close the landing gear, Landing Gear Down, Speed-braker Arming, Turn on the Reverse Thrust and Retract the Reverse Thrust) are no difference in two tasks. This quantitative validates the cognitive that

compared with the normal situation, the case of a large crosswind need to consider more factors, the technology is more complicated [24]. At the same time, from the point of view of engineering application of aviation safety, for the first time, the coding scheme was developed for the pilot's behavior, providing new ideas for studying pilots' specific behaviors, which laid a certain foundation for the future application of Observer XT(12.0) software to study the pilot's behavior. However, this paper only uses Observer XT (12.0) to conduct preliminary research on pilots'. The study on the pilot's behavior still needs to be further deepened, mainly in terms of quantitative analysis, which makes it better applied to the study of human factors of aviation safety.

4 Conclusion

Pilot's operational performance has a direct impact on flight safety. Using Observer XT (12.0) to study the pilot's cockpit's operational behavior can reduce interference with the pilot's current activity, ensure the objectivity and accuracy of the analysis and, to the maximum extent, ensure that the behavior we want to observe is not impacted of additional variables in the environment. More than 20 years after its birth, this software has been widely used in human factors research [10] and human-computer interaction research [16]. In this paper, the study of pilots' behavior is based on reading the "Aircraft Flight Manual" [6] and watching the "Eye of the Flight" and other methods to design pilots cockpit behavior encoding scheme. Based on the analysis of operational behavior data, we can get the following conclusions: Our developed scheme for coding pilot actions provides a useful way to quantify the pilots' specific behaviors in the future. However, the coding scheme does not include all pilot actions and needs to be further improved; By building an experimental platform suitable for observing pilots' behavior, the typical behaviors of pilots were analyzed by Observer XT 12.0, and the similarities and differences in pilot behaviors under two tasks were compared. This provides a method for analyzing the behavior of pilots in the future.

References

1. Ban, Y.K.: Aviation Accidents and Human Factors. China Civil Aviation Press (Chinese), Beijing, pp. 30–38 (2002)
2. Bonomalenko, B.A., Lapa, B.B.: Flight Psychology, pp. 40–43 (1988)
3. Chen, H., Wang, G.: Pilot control behavior analysis using cutoff frequency and power frequency for a civil transport aircraft encountering turbulence based on flight Simulation. *Procedia Eng.* **80**, 424–430 (2014)
4. Chen, N.T., Tan, X.: Experimental design of pilots' operational behavior analysis based on electromyography and skin temperature detection. *Exp. Technol. Manag.* **32**(11), 202–205 (2015). (Chinese)
5. Chen, M.L.: Comparison on the teaching behaviors of expert and novice aerobics teacher. Doctoral dissertation, Wuhan Institute of Physical Education (Chinses) (2013)
6. Federal Aviation Administration: Aircraft Flight Manual. Shanghai Jiao Tong University Press (Chinese) (2010)

7. He, W., Ke, S.H., Wu, X.B., Li, X.L.: Relationship between crew behavior, time margin and flight safety. *J. Saf. Environ.* **3**(2), 16–18 (2003). (Chinese)
8. Hayashi, K., Suzuki, S., Uemura, T.: Analysis of human pilot behavior at landing with neural network. *J. Jpn. Soc. Aeronaut. Space Sci.* **49**(564), 21–26 (2001)
9. Hillard, D., Manavoglu, E., Raghavan, H., Leggetter, C., Iyer, R.: The sum of its parts: reducing sparsity in click estimation with query segments. *Inf. Retr.* **14**(3), 315–336 (2011)
10. Heffelaar, T., Kuipers, J., Andersson, J., Wiertz, L., Noldus, L.P.J.J.: Easy to use driving behavior analysis using drive lab. In: Stephanidis, C. (ed.) *HCI 2014. CCIS*, vol. 434, pp. 330–334. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07857-1_58
11. Keane, T.: Combat modeling with partial differential equations. *Appl. Math. Model.* **35**(6), 2723–2735 (2011)
12. Liu, R., Liu, Z.H.: Research on psychology determination system of flight behavior observation. In: *Automation and Instrumentation* (2017)
13. Liu, S.Q., Sun, Y.C.: Research on crew workload base on BP neural network. *Aircr. Des.* (2), 63–66 (2014). (Chinese)
14. Luo, F., She, L., Gu, B.C.: The psychological analysis on the behavior mistakes of the civil aviation crews. *J. Wuhan Univ. Technol. (Transp. Sci. Eng.)* **26**(2), 191–194 (2002). (Chinese)
15. Liu, Z.H.: Upper extremity behavior for the pilot. Doctoral dissertation, Shanghai Jiao Tong University (Chinese) (2012)
16. Li, P.F.: Research on indices and analysis of driving behavior. Doctoral dissertation, Jilin University (Chinese) (2010)
17. Liu, Y.J., Sun, Y.C.: Aircraft control process modeling and ergonomics analysis based on CPN. *Aeronaut. Comput. Tech.* **47**(1), 69–73 (2017). (Chinese)
18. *NASA Vision 2050—An Integrated National Transportation* (2011)
19. Smith, R.E., Dike, B.A., Ravichandran, B., El-Fallah, A., Mehra, R.K.: Discovering novel fighter combat maneuvers: simulating test pilot creativity. In: *Creative Evolutionary Systems*. Morgan Kaufmann Publishers Inc (2002)
20. Sun, R.S., Xiao, Y.B.: Research on indicating structure for operation characteristic of civil aviation pilots based on QAR data. *J. Saf. Sci. Technol.* **8**(11), 49–54 (2012). (Chinese)
21. Wu, L., Wang, X.F.: Mathematics analysis method research for dependence of flight crew operation behavior. *Sci. Technol. Innov. Her.* **12**(26), 5–9 (2015). (Chinese)
22. Xue, H.J., Pang, J.F., Luan, Y.C., Li, L.: Cockpit pilot cognitive behavioral integration simulation modeling. *Comput. Eng. Appl.* **49**(23), 266–270 (2013). (Chinese)
23. Yin, Y.F., Guan, H.C., Zeng, Y.F., Sun, T.H.: Pilot dynamic behavioral evaluation method. *J. Chongqing Univ.: Nat. Sci. Ed.* **36**(6), 154–160 (2013). (Chinese)
24. Zhang, Z.P.: Take-off and landing skills under big crosswind. *Saf. Secur.* **6**, 54–55 (2014). (Chinese)
25. Zimmerman, P.H., Bolhuis, J.E., Willemsen, A., Meyer, E.S., Noldus, L.P.: The observer xt: a tool for the integration and synchronization of multimodal signals. *Behav. Res. Methods* **41**(3), 731–735 (2009)



Tablet-Based Information System for Commercial Aircraft: Onboard Context-Sensitive Information System (OCSIS)

Wei Tan^{1(✉)} and Guy A. Boy^{2(✉)}

¹ School of Flight Technology, Civil Aviation University of China, Tianjin 300300, China
weitan2011@outlook.com

² ESTIA/Air and Space Academy, 64210 Bidart, France
guy.andre.boy@gmail.com

Abstract. Pilots currently use paper-based documentation and electronic systems to help them perform procedures to ensure safety, efficiency and comfort on commercial aircrafts. Management of interconnections among paper-based operational documents can be a challenge for pilots, especially when time pressure is high in normal, abnormal, and emergency situations. This dissertation is a contribution to the design of an Onboard Context-Sensitive Information System (OCSIS), which was developed on a tablet. The claim is that the use of contextual information facilitates access to appropriate operational content at the right time either automatically or on demand. OCSIS was tested using human-in-the-loop simulations that involved professional pilots in the Airbus 320 cockpit simulator. First results are encouraging that show OCSIS can be usable and useful for operational information access. More specifically, context-sensitivity contributes to simplify this access (i.e., appropriate operational information is provided at the right time in the right format. In addition, OCSIS provides other features that paper-based documents do not have, such as procedure execution status after an interruption. Also, the fact that several calculations are automatically done by OCSIS tends to decrease the pilot's task demand.

Keywords: Commercial aircraft · Onboard Information System
Human-centered design · Tangible interactive system · Avionics · Context

1 Introduction

An airplane consists of a number of mechanical and computerized systems. An airplane cannot stop or brake in the air, and fuel is consumed during the entire flight. Consequently, flight time is limited. Flight crewmembers have all the capacities and limitations of any human being; they can be qualified as human operators. They typically collaborate, communicate, and cooperate with each other to execute flight tasks.

Procedures execution is a major safety factor. Until now, specific pilot roles have been supported by paper-based documentation in both operations and training. It is a pilot's job to make decisions, act, communicate, cooperate, and coordinate operationally, with procedures established in operational documentation developed through

airline policy and governmental regulation. All the procedures and information can be found from documents. The onboard documents can be categorized into four kinds of documents: flying documents, which are related to all flight operations; systems documents, which include systems' theory, principles, and controls; navigation documents, which are the charts that pilots use on the flight deck; and performance documents, which provide operational data for all flight phases such as takeoff, landing, and go-around [1]. Actions performed by flight crewmembers in the cockpit must adhere to procedures in context. Onboard paper-based operational documents barely provide context-sensitive information. Therefore, context has to be handled by pilots.

However, onboard paper-based documents are not the only resources that pilots have in the cockpit. Several other onboard systems can enhance the pilot's awareness of aircraft status (i.e., aircraft states). They can provide very comprehensive information on the state of the aircraft in an integrated way. Taking Airbus Electronic Centralized Aircraft Monitor (ECAM) as an example, it provides actions together with corresponding flight parameters to pilots who have to deal with related malfunctions. It provides steps to handle failures for a large number of situations [2].

2 State of Art

2.1 Tablet-Based Systems Onboard

Pilots are familiar with paper-based manuals, which are easy to use, tag, mark, and retain, even though they are heavy and difficult to carry. Nobody can permanently remember all procedures and technical knowledge, particularly, under time pressure. Now many applications on tablets that contain paper charts information are available. Moreover, Boeing introduced a tablet-based version of the paper Quick Reference Handbook (QRH) used by flight crews in 2013 [3]. The Interactive QRH offers advanced navigation and search capabilities to enable the pilot to easily find the proper checklist. It also simplifies non-normal checklist use, especially for those checklists in which the correct condition must be selected from two or more choices. In line with technological innovation, Airbus has developed iPad EFB solutions to provide airlines with an alternative to PC operating system EFB devices. With "FlySmart with Airbus" applications on iPad, pilots will be able to compute performance calculations and also consult Airbus Flight Operations Manuals from a light hand-held device [4, 5].

Even though Electronic Flight Bag (EFB) and Onboard Information Systems (OIS) are advanced tools to assist pilots' work, operational documents are still in their original format and arrangement. Not all abnormal procedures are available on the ECAM, nor do all types of aircrafts have ECAM or some other electronic systems to process and display procedures. Unlike paper, computer support enables easy contextualization of provided information. The Onboard Context-Sensitive Information System (OCSIS) is introduced into commercial aircraft cockpit to make the flight safe, efficient, and comfortable by providing assistance in normal and abnormal situations and enhanced capabilities of interaction with other onboard systems.

2.2 Human-Centered Design Approach

“A human-in-the-loop (HITL) simulation is a modeling framework that requires human interaction. This approach is typically called participatory design. The emergence of HITLS technologies, therefore, enables researchers and practitioners to investigate the complexity of human-involved interactions from a holistic, systems perspective [6] ”. As shown in Fig. 1, the model typically represents reality in a simplified way. It proposes important elements and their relevant interconnections in an appropriate, orchestrated manner. The simulation represents the interaction that brings the model to life, which can be used to improve understanding of interactions among different elements that the model implements. It is also used to improve the model itself and eventually modify it [7]. HITLS is used early on during the design phase.

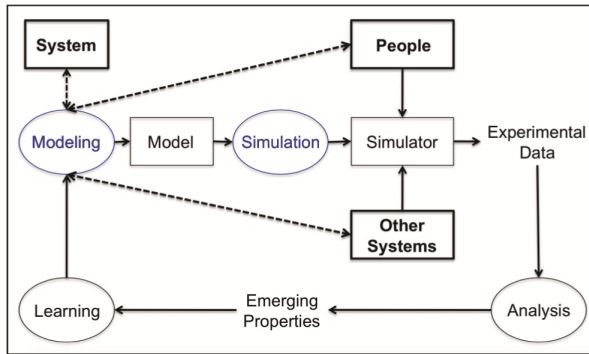


Fig. 1. Human-centered design approach [7]

Consequently, Human-Centered Design (HCD) has been used to incrementally improve OCSIS toward an acceptable mature version (i.e., incremental prototype development, test, and modification) [8]. Modeling OCSIS requires pilots’ involvement, and interaction with other onboard systems. The process can be run on a flight simulator, which in turn produces experimental data that could be used to improve OCSIS.

2.3 Context-Sensitive Procedures

“Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves [9] ”. Pilots must accomplish flight tasks (e.g., cockpit preparation, takeoff, approach, and landing procedures) within the appropriate context depend on flight phases. We use Interaction Blocks to represent, implement, and handle context-sensitive procedures (see Fig. 2). An interaction block is defined by: a set of actions; and a situation pattern that includes triggering preconditions and a context pattern; and post-conditions that include a goal and abnormal conditions [10]. This procedural knowledge representation was developed during the 1990s to represent operational procedures in aircraft cockpits and

led to deeper investigations on context representation also. It is therefore very appropriate for context-sensitive procedural information representation.

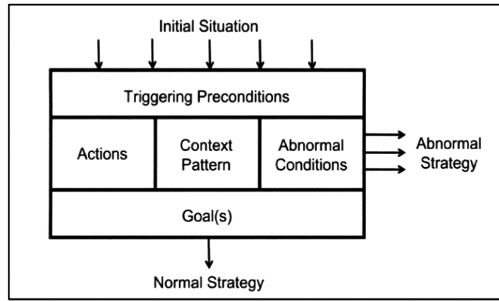


Fig. 2. Interaction block representation [10]

3 Design of OCSIS Prototype

OCSIS is currently programmed using Objective-C, an object-oriented language available on Apple’s hardware, on Xcode, the Integrated Development Environment for Objective-C. The first prototype of OCSIS is applying A320 procedures and references. Once the application starts on the iPad, a Welcome page is displayed. The default/initial page displays procedures and actions that crews need to perform or have performed. A menu is provided to select other tabs at the top of this page (see Fig. 3).

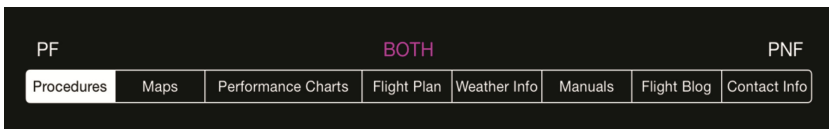


Fig. 3. OCSIS’s menu

At the end of the 1990s, EURISCO and Airbus carried out a study with 60 commercial pilots on how electronic documentation could be structured [11]. Results showed that it could be best structured into three information levels. We adopted this re-organized structure for OCSIS (see Fig. 4):

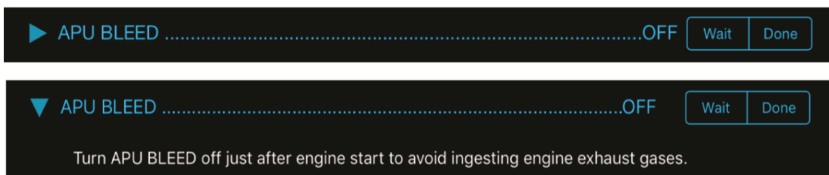


Fig. 4. Level 1 and Level 2 information of APU Bleed set

Level 1: Need to know or safety-critical information that the pilot needs to have immediately.

Level 2: Nice to know or short explanations of Level 1.

Level 3: Understand technical knowledge on systems’ principles and trouble-shooting.

In order to keep consistency of Airbus’s philosophy and make all the actions are easy to be understand, Dynamic Color System (DCS) is designed to enhance the pilot’s perception and comprehension of the current situation., as shown in Table 1. Different colors stand for different meanings that provide the pilot with a direct and swift status of procedures. The items marked with “*” are updated after a series of formative evaluations.

Table 1. Color codes

| Color | Representation |
|---------|---|
| Cyan | Actions to be performed |
| Green | Actions performed Marked as performed |
| Amber | Postponed actions or checks *The title of abnormal procedures Cautions |
| Red | *The title of emergency procedures Warnings |
| White | Notes More information for actions *The title of flight phase *Normal checklists |
| Magenta | Restrictions or constraints |
| Grey | *Not applicable for current context |

“Ready to do” actions are in cyan. Once the action is completed and OCSIS can access the related parameters’ status, it automatically becomes “green.” In the current version of OCSIS, this kind of automation is done for only the a few parameters, which can be detected (i.e., colors change automatically). For all other parameters, the pilot marks them “DONE” manually (i.e., the action line becomes green). The pilot can also postpone an action by selecting the “WAIT” button, and then the action line becomes amber (see Fig. 5).



Fig. 5. Cyan, green, amber, and white for action status (Color figure online)

“Scenarios of human-computer interaction help us to understand and to create computer systems and applications as artifacts of human activity — as things to learn from, as tools to use in one’s work, as media for interacting with other people [12]”. The current version of OCSIS includes several normal procedures and two abnormal scenarios. “Initial Approach” and “Final approach” are the scenarios that we choose for normal procedures. “Fuel Leak” and “Flaps Locked” are the scenarios that we choose for abnormal procedures. Context patterns trigger procedures in real-time both in normal and abnormal situations. In an abnormal situation such as “Flaps Locked,” OCSIS will immediately inform the pilot about this malfunction by displaying a pop-up information window (see Fig. 6). Pilots can become aware of the problem through the pop-up window and start following actions. If they choose to do it later, a reminder line (see Fig. 7) will be displayed at the bottom of the interface, which directs to additional “Flaps Locked” procedures.

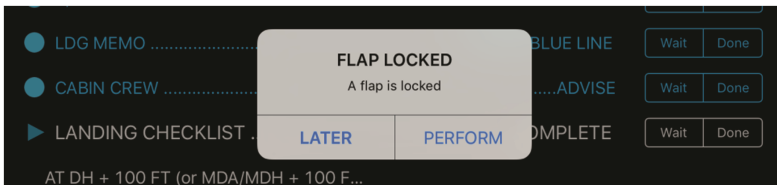


Fig. 6. “Flaps Locked” triggering

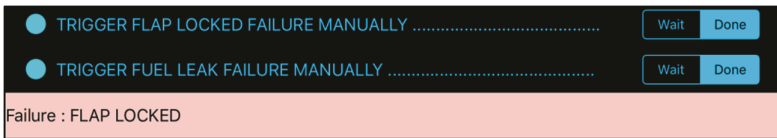


Fig. 7. “Flaps Locked” reminder

4 Formative Evaluation

OCSIS testing complies with the four key types of human factors defined by Chanda and Mongold [13]: (1) Usability of hardware user interface (i.e., we used questionnaires to find out about the location of the OCSIS iPad in the cockpit); (2) Usability of software user interface (i.e., we used questionnaires to get feedback from pilots on usability and usefulness of interface items); (3) Integration of hardware and software with existing cockpit systems (i.e., pilots were asked to provide their opinions on the operational integration of OCSIS in the cockpit); (4) Design of training/procedures for OCSIS (i.e., we designed normal and abnormal procedures for second and third test).

The first testing was carried out with four pilots. They were involved in two sessions. The first session consisted of performing all required procedures both in normal and abnormal scenarios using paper-based manuals. The second session consisted of performing the same procedures using OCSIS. There was an interval of a few days between the two sessions for each person.

Pilot participants provided excellent feedback on actions to use OCSIS, look and feel, and other usability criteria. The results showed that every pilot understood how to use OCSIS. They all reported that OCSIS was easy to hold and use. Pilots provided feedback on information icons, color, and size, which was used to improve user interaction with OCSIS. Results showed that all pilots, except one (who said that size of the items was not big enough), felt comfortable with OCSIS information display and interaction.

Based on pilots' feedbacks we made improvements on part of the icons' color and buttons' function to OCSIS, also changed flight phase titles' size, position and function to remove the ambiguity. A grey color code was added to the system representing the "Not Applicable" case to assist pilots' decision-making. There were procedures that are embedded into other procedures that cause recursivity issues, e.g., the Engine 1(2) Relight procedure is embedded into the "Fuel Leak" procedure. During this phase of testing, we simply added the content of the Engine 1(2) Relight procedure into the "Fuel Leak" part, but we subsequently developed a generic hyperlinked iBlock system within OCSIS that enables to automatically put into an operational sequence.

An integration survey was administered to the pilots. Results showed that all pilots prefer that the iPad be fixed in the cockpit rather than available as a handheld device. Four options were suggested (see Fig. 8): (1) Next to side-stick for each pilot; (2) On the windshield with a flexible arm; (3) In the pedestal or on the instrument panel as a unit; (4) Inside a box at the side of the pedestal.



Fig. 8. Options of iPad's location

The second testing was carried out at a flight training center in China. Twenty-two A320 pilots, including eleven captains and eleven first officers, participated in the testing, performing as aircrew in A320 simulators. We used the same route and protocols as the first testing. During the testing, we observed in the "Fuel Leak" scenario pilots easily established the failure and excluded the irrelevant procedures (e.g., "Fuel imbalance" procedure). In the "Flaps Locked" scenario, OCSIS reduced the chance of wrong calculation of landing distance and approach speed. Regarding OCSIS look and feel,

pilots were asked to evaluate OCSIS usability in terms of colors, interactive buttons, and other devices available on the OCSIS iPad. Results showed that three pilots had difficulty during training to select buttons.

Two pilots thought OCSIS was not reactive enough; this was due to the fact that data-link was not available on the simulator at that time. We corrected that. Pilot-OCSIS interaction was evaluated using pilots’ feedback on information icons, color, and size. The results showed that pilot-OCSIS interaction is satisfactory in general.

Based on pilots’ feedback, we made improvements to OCSIS after second testing: “Engine Shutdown” and “Engine Relight” procedures were included in the “Fuel Leak” scenario. Pilots may shut down engine to check if fuel leak is from engine or somewhere else. When pilots move the thrust lever to idle position, the low-speed rotor (N1) of engine is going to be 0, which triggers the “Engine Failure” procedure both on ECAM and OCSIS that increases redundancy. Pilots should perform ECAM actions first and then come back to OCSIS to complete additional actions of “Engine Shutdown” procedure. This being done, pilots can continue executing the “Fuel Leak” procedure, and pilots had to remember to go back procedure. Pilots may decide to relight the engine if they check and discover that the fuel leak is not coming from the engine but from elsewhere. The “Engine Relight” procedure is provided on the same page (see Fig. 9).

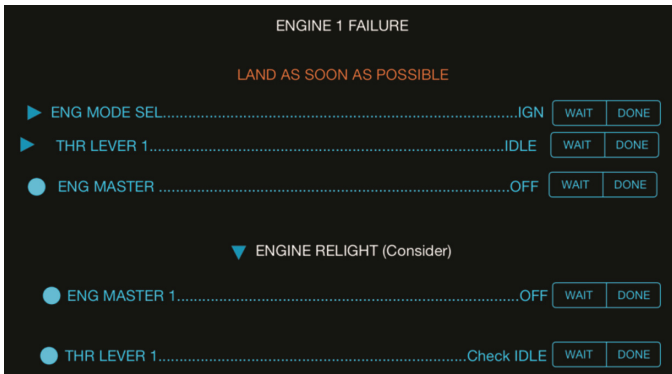


Fig. 9. Engine relight procedure.

Fourteen pilots preferred the iPad located close to each side-stick (Position 1 in Fig. 12). Eight pilots preferred the iPad installed in the middle of instrument panel (Position 2 in Fig. 8). We chose Position 1 as the test location of OCSIS in the third testing. A flexible arm was set up to hold the iPad near the side-stick on the right side of the simulator.

The third testing was held with six pilots participated, and three additional pilots took the Nielsen’s ten Usability Heuristics survey [14], using the same protocols and timelines as first and second testing. The results of user-system interaction questionnaires showed that every pilot understood and was satisfied with using OCSIS and with OCSIS’s user interaction, as well as the location of OCSIS in the cockpit. Pilots were required to assess OCSIS look and feel by evaluating display usability in terms of color, buttons, and other OCSIS devices. The results show that all pilots had no trouble to

understand the system. Pilot-OCSIS interaction was assessed on the basis of pilots’ feedback on information icons, color, and size.

Based on Nielsen’s ten Usability Heuristics survey [14] which helped us prioritize issues with respect to those that users found critical to those that may not be critical, we made the following improvements to OCSIS after third testing mainly on design decision phase:

1. Quick maneuverability to specific procedures/menus instead of scrolling (see Figs. 10 and 11). It is possible to fit each flight phase on a single page, and pilots can move left and right to review procedures for other flight phases. In the meantime, a menu to select a particular flight phase can be set at the top of the screen. Pilots should be free to move through menus; information flow progress is saved on each page.



Fig. 10. Procedure’s headings before improvement

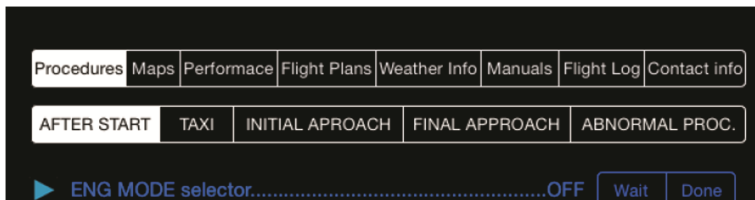


Fig. 11. Procedure’s headings after improvement

2. Consistency is a key aspect in usability engineering. For example, solid clickable boxes such as “Check-all” bars are inconsistent with the usual cockpit format (e.g., the background of other clickable boxes is black and characters are blue, and as shown on Fig. 12. The solid clickable box format of Check-all bar is the opposite. Figure 13 is showing the improvement on this point).

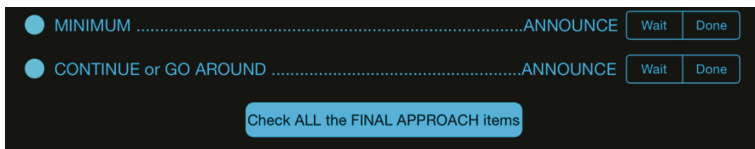


Fig. 12. Checklist’s icon before improvement (Color figure online)

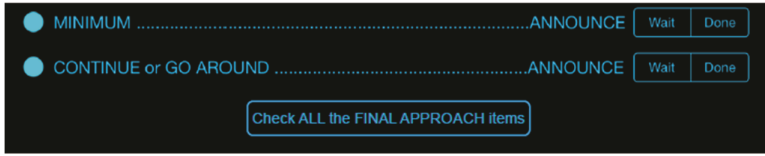


Fig. 13. Checklist’s icon after improvement

- 3. Caution or warning messages as well as titles of abnormal procedures should be color coded. Moreover, abnormal procedure headings are not very obvious for every section/page/title of an abnormal procedure. It should be larger (e.g., white “Fuel Leak” title in Fig. 14. It is modified to be amber in Fig. 15).



Fig. 14. The title of abnormal procedure before improvement (Color figure online)

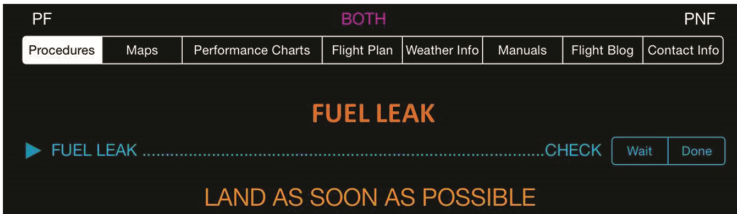


Fig. 15. The title of abnormal procedure after improvement (Color figure online)

5 Discussion

OCSIS was first designed to be used as a tangible interactive system (TIS) onboard a commercial aircraft [15]. Onboard paper-based documentation has been used from the beginning of aviation history and is tangible for pilots to use. Tablets and some applications (e.g., Jeppesen Mobile FliteDeck) have been authorized to be used on the flight deck by the FAA. Consequently, we considered that tablets are tangible objects that can support OCSIS software. This is physical tangibility, but figurative tangibility should be tested [16]. Figurative tangibility in this case means for pilots to keep correct cognition using OCSIS. The testing studies were conducted to provide a first set of methods

and tools with this figurative tangibility assessment. And it is possible for us to extend this assessment. The testing studies that were performed showed that this hypothesis was confirmed on a fully equipped cockpit simulator in realistic flight operations scenarios with professional pilots. It is obvious that more iteration is needed and will be implemented in the near future to get a mature version of OCSIS.

This paper provides a first iteration of participatory design of OCSIS. More generally, it shows the shift from the traditional automation approach, where additional software was added to the cockpit and induced some kinds of rigidity that sometimes resulted in unexpected situations, to TIS design, where tangibility has to be tested using situation awareness models and criteria [17]. Of course, the concept of tangibility is more complex and will require more investigation.

6 Conclusion

This research and HCD effort are based on both participatory design and agile development (i.e., at the end of each phase, the system is testable in an HITLS environment). This is now typical for the design and development of tangible interactive objects [7], and more generally tangible interactive systems (TISs) [16], where the problem is no longer automation but the search for tangibility. Modelling and simulation are very useful to explore possibilities and drawbacks of these TISs. The quality of both simulation capabilities and pilot participants is crucial [8]. If the issues traditionally raised by human factors and ergonomics specialists when engineering work is done are now posed at the beginning of the design phase in a virtual world (i.e., virtual engineering is part of HCD), new kinds of questions would emerge from this practice, that is, tangibility [18].

OCSIS is a comprehensive system that aims to make flight of commercial aircraft safe, efficient, and comfortable. The multiple usability evaluations and user-centered assessments performed on the system can discover the maximum number of issues. First designs of OCSIS were based on our creativity process, in the sense of synthesis and integration, on previous expertise and experience in the commercial aviation domain, more specifically, work done by Blomberg, Speyer and Boy [11] on the three layers of electronic operations documentation and Ramu's [19] dissertation work on onboard context-sensitive information systems. This work is typically based on human-computer interaction (HCI), hypertext, and context-sensitive information systems work. Usability testing brought us a series of usability issues that helped us to concretely improve OCSIS. Although not all the solutions could be addressed due to time, these can be addressed in further work.

Acknowledgments. We thank Sebastien Boulnois and Anouk Mirale, who developed a fair amount of OCSIS software, and Varun Korgaonkar who designed the usability questionnaires for the third testing. Thanks to our HCDi colleagues, Jarrett Clark, provided software development support for this project. Dr. Alexandre Lucas Stephane, helped me with the design of testing and the analysis of experimental results, and Dr. Ondrej Doule, who helped me with the design theory and methods. Thanks to all who participated in the various tests and provided me with their time and great feedback to improve OCSIS. Thanks to Dr. Barbara K. Burian, Dr. Divya Chandra, Dr.

Christophe Kolski, and Dr. Scott Winter, who provided me with feedback, suggestions, and comments on this project.

References

1. Tan, W.: From commercial aircraft operational procedures to an onboard context-sensitive information system. In: Proceedings of the HCI-Aero Conference, Santa Clara, CA (2014)
2. Airbus Company: Airbus 380 Flight Crew Operating Manual
3. Boeing. www.boeing.com/boeingedge/aeromagazine. AERO, Issue 49_QTR_01 (2013)
4. Airbus press release: combining airbus' EFB content with the world's most versatile mobile device (2012). <http://www.airbus.com/presscentre/pressreleases/press-release-detail/detail/airbus-offers-ipad-electronic-flight-bag-solution/>
5. EASA EFB Evaluation Report: European Aviation Safety Agency Electronic Flight Bag (EFB) Evaluation Report. Airbus FlySmart with Airbus for iPad – V2. 12 (2013)
6. Narayanan, S., Rothrock, L.: Human-in-the-Loop Simulations: Methods and Practice. Springer Science & Business Media, Berlin (2011). <https://doi.org/10.1007/978-0-85729-883-6>. ISBN 0857298836
7. Boy, G.A.: From automation to tangible interactive objects. *Ann. Rev. Control* (2014). <https://doi.org/10.1016/j.arcontrol.2014.03.001>. Elsevier (1367–5788)
8. Boulnois, S., Tan, W., Boy, G.A.: The Onboard Context-Sensitive Information System for Commercial Aircraft. In: Proceeding 19th Triennial Congress of the IEA, Melbourne, Australia (2015)
9. Dey, A.K.: Understanding and using context. *Pers. Ubiquit. Comput.* **5**(1), 4–7 (2001)
10. Boy, G.A.: Cognitive Function Analysis, vol. 2. Greenwood Publishing Group, Westport (1998)
11. Blomberg, R., Boy, G.A., Speyer, J.J.: Information needs for flight operations: human-centered structuring of flight operations knowledge. In: Proceedings of HCI-Aero 2000, pp. 45–50. Cépaduès-Éditions, Toulouse (2000)
12. Carroll, J.M.: Five reasons for scenario-based design. *Interact. Comput.* **13**(1), 43–60 (2000). ISSN 0953-5438
13. Chandra, D.C., Mangold, S.J.: Human factors considerations for the design and evaluation of electronic flight bags. In: Proceedings of 19th Digital Avionics Systems Conference, ISBN 9780780363953, vol. 2, pp. 5A1/1–5A1/7 (2000)
14. Nielsen, J.: Usability Engineering. Academic Press, Boston (1993). ISBN 0-12-518405-0
15. Tan, W., Boy, G.A.: Iterative designs of onboard context-sensitive information system (OCSIS). In: Proceedings of the HCI-Aero Conference, Paris, France (2016)
16. Boy, G.A.: Tangible Interactive Systems - Grasping the World with Computers. Springer, Heidelberg (2016). <https://doi.org/10.1007/978-3-319-30270-6>
17. Tan, W.: Contribution to the onboard context-sensitive information system (OCSIS) of commercial aircraft. Ph.D. Dissertation. Florida Institute of Technology (2015)
18. Tan, W., Boy, G.A.: Iterative designs of onboard context-sensitive information system (OCSIS) for commercial aircrafts. *J. Transp. Inf. Saf.* **34**(4), 70–77 (2016). ISSN 1674-4861
19. Ramu, J-P.: Contextual operational documentation effectiveness on airline pilots' performance. Ph.D. Dissertation. ISAE, Toulouse (2008)



Modeling and Simulating Astronaut's Performance in a Three-Level Architecture

Chunhui Wang¹, Shanguang Chen^{1,2(✉)}, Yuqing Liu¹,
Dongmei Wang³, Shoupeng Huang¹, and Yu Tian¹

¹ National Key Laboratory of Human Factors Engineering,
China Astronaut Research and Training Center, Beijing 100094, China
chunhui_89@163.com, shanguang_chen@126.com,
huangshoupeng2005@163.com, clara@163.com,
cctian@126.com

² China Manned Space Program, Beijing 100080, China

³ School of Mechanical Engineering, Shanghai Jiao Tong University,
Shanghai 200240, China
dmwang@sjtu.edu.cn

Abstract. Astronaut's capabilities to successfully complete specific tasks during missions is of vital importance for spaceflight. While in spaceflight, astronauts are exposed to numerous stressors, such as microgravity, confinement and radiation, all of which may impair human capabilities. So it is crucial to get a better understanding of astronauts' capabilities and to better predict their task performance during long-term spaceflights. Computer models can be used to learn from and even predict human performance, which can enhance early evaluation of system designs, and reduce the time cycle and costs of system development. To support modeling and simulation of astronaut's performance in specific physical and cognitive tasks during spaceflight, we established the Astronaut Modeling and Simulation System (AMSS), which is the first integrated modeling and simulation platform for human-system integration design faced to long-duration manned space missions in China. A three-level model architecture has been proposed, which consists of the human characteristic models, the behavioral models (cognitive and biomechanical) and the performance evaluation models. The multiple models are integrated in AMSS. The ability to visualize the virtual environment of space vehicle, the virtual astronaut, the operator's performance and task processes makes AMSS a user-friendly platform. Models in AMSS have been preliminarily validated by experimental data. AMSS has been used to perform the quantitative evaluation of the human-machine interface designs of China's space lab and the on-going space station missions, and will be applied to the human-system integration design in China's future space missions.

Keywords: Astronaut · Performance · Modeling and simulation

1 Introduction

It is well known that astronauts play a very important role in manned space missions. In order to ensure the successful implementation of manned space missions, it is essential to maintain and improve the capabilities and task performance of astronaut in space [1].

However, in spaceflight, astronauts are exposed to numerous stressors, such as microgravity, confinement, and radiation, all of which may impair human capabilities, and increase task risks [2]. So the mismatch between crew capabilities and task demands has been listed as a major risk that astronauts may encounter in space in the human research roadmap published by the National Aeronautics and Space Administration (NASA) [3]. If tasks prove too difficult for the operators, whether as a result of inadequate design of man-machine interfaces or task schedules, or capability decline in spaceflight, the work efficiency of the space crew may decrease and the likelihood of mission failure increases. So it is crucial to get a better understanding of astronauts' capabilities and their task performance during spaceflights.

Computational simulations, especially those combined with human performance models (HPMs), have been used more and more frequently in space explorations and aviation to analyze human performance and identify potential human-system errors, due to the cost and efficiency advantages they have over traditional human-in-the-loop tests [4, 5]. HPM simulations can be used early in the development process of a product or system, thus are highly applicable to human-system integration design of space tasks and systems. Specially, some aspects of the spaceflight environmental factors, such as the weightlessness, are difficult to simulate on ground, while HPM may offer an effective mean to simulate human behaviors and performance in such environment, thereby providing useful guidance for human-system integration design.

There are generally two kinds of HPMs that are commonly used in the space and aviation industry: physical models (models of human anthropometry, models of biomechanics), and cognitive models built from empirical research and theories of human cognitive processes; the tasks investigated can also be categorized into two types: physical tasks and cognitive tasks. A lot of software tools have been developed to support human performance modeling. For example, Jack and DELMIA, which consist of models of human anthropometry, have been used to analyze the human reach envelope, human visual field and space interference problem [6]. Anybody, OpenSim and Visual3D, which consist of models of human biomechanics, have been used for force analysis of human body (including joint torque, muscle force, etc.) in physical task [7]. ACT-R and Soar, which consist of human cognitive models, have been used to simulate decision making processes in cognitive tasks [8]. The Man-machine Integration Design and Analysis System (MIDAS), developed by NASA, combines graphical equipment prototyping, a dynamic simulation, and human performance modeling in an integrated platform [9]. Cognitive models are core elements of MIDAS. The aim of MIDAS is to reduce design cycle time, support quantitative predictions of human-system effectiveness, and improve the design of crew stations and their associated operating procedures.

However, most of models and platforms (except MIDAS) are not suitable for the simulation analysis of astronaut performance in spaceflight tasks. They lack systematic considerations of the space environments, astronaut's characteristics (especially Chinese astronauts' characteristics in space), and the space task demands, which may all affect astronaut performance in space. As far as we concern, to model Chinese astronauts' performance in space, especially their performance during long-term spaceflight,

cognitive models and physical models should be combined to produce an integrated platform, and a comprehensive model architecture and platform need to be proposed and built.

2 The Architecture and Models of AMSS

2.1 The Overall Model Architecture

The objective of the current paper is to develop a simulation platform to predict human performance in spaceflight. There are three key aspects about the astronaut's performance modeling: (1) how to use human biomechanical and cognitive parameters to characterize the individual differences, and how these parameters change in spaceflight environmental; (2) how to describe the task features and human-system interaction, such as task type, task procedure, man-machine interfaces, and demands of the task; (3) how to evaluate the human-system effectiveness accurately and quantitatively. Those three aspects were particularly emphasized in our architecture design, and a three-level model architecture, which consists of human characteristic models, behavioral models, and performance evaluation models, were proposed accordingly, as show in Fig. 1.

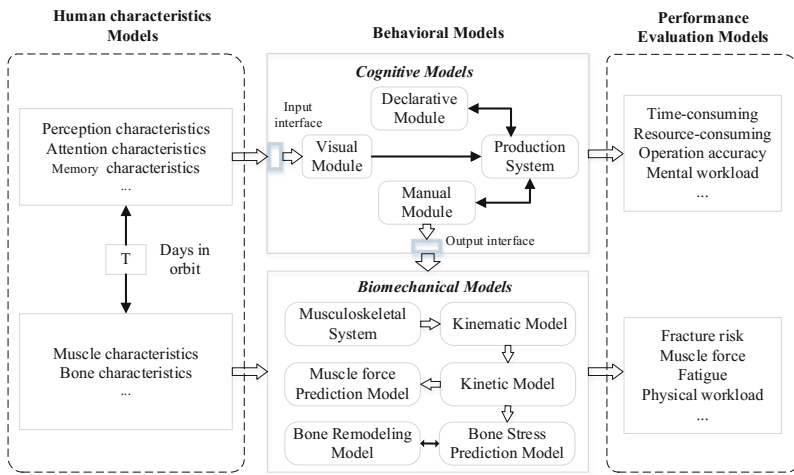


Fig. 1. Three-level model architecture for astronaut performance modeling

In the three-level model architecture, the first level contains a database of astronauts' characteristics and some models which can describe the changes of astronauts' capabilities during spaceflight. The second level contains two kinds of behavioral models: biomechanical models and cognitive models. The biomechanical models are used to calculate joint forces, muscle forces, bone stress, bone mineral density in space tasks. Cognitive models are used to simulate the cognitive processes in performing

space tasks. On the third level, we build a set of performance evaluation models, so we can predict astronaut's task completion time, mental workload and physical workload, bone fracture risk, etc.

In this model architecture, the characteristics models mainly reflect the influence of spaceflight environmental on astronaut; the behavioral models mainly reflect the influence of task characteristics on astronauts' performance. Together with the performance evaluation function, the main human factors concerns are considered in the model architecture.

2.2 Human Characteristics Models

The term "human characteristics" in this paper refers to human biomechanical parameters (such as maximum muscle strength, maximum joint torque), and human perceptual and cognitive characteristics. As many human characteristics will change, it is very important to consider the change regularities of human characteristics in spaceflight when we design a space system. The change regularities of human characteristics in spaceflight are defined as human characteristics models in this paper. The human characteristics models serve as input into behavioral models, and thereby affect the task performance, which reflects the influence of spaceflight environment on human performance. Human characteristics data were collected in spaceflight, or simulated space environment, such as experiments during parabolic flights, the head-down-bed-rest experiment, isolation experiments.

Biomechanical Characteristics

In long duration spaceflight, muscular atrophy and bone loss will have significant impacts on astronaut's biomechanical characteristics, eventually affect astronaut health and performance. Among the many biomechanical characteristics, the maximum muscle strength, maximum joint torque and the bone density, as well as their changes in spaceflight and simulated weightlessness, are primarily investigated.

The maximum muscle strength are measured in parabolic flight experiments and Shenzhou-10 and Shenzhou-11 missions. As the measurement of these parameters in the parabolic flight experiments and spaceflight missions have many restrictions, such as a short test time in parabolic flights, we also test 16 subjects' maximum muscle strength, muscle circumference of the shins and thigh, maximum joint torque, bone density in a 45d -6°head-down bed rest experiment. All these data help us to build the models which can reflect the change regularity of the maximum muscle strength, muscle circumference, maximum joint torque [10]. Bone density remodeling is a complicated biological activity, which must be represented in models of nonlinear characteristics. A bone remodeling control equation was used to describe the relationship between density changes and mechanical loads, which can predict bone density in spaceflight [11]. Additionally, the quantitative relationship among bone elastic modulus and bone mineral density was obtained by testing the bone elastic modulus of 3 corpses [12].

Cognitive Characteristics

By cognitive task analysis of the space tasks, a battery of cognitive characteristics that may affect task performance were extracted [12], including perception characteristics

such as reaction times, speed perception, time perception, the working memory and prospective memory, spatial ability, fine motor control, attentional control, emotional characteristics, risk decision-making characteristics, etc. Experiments were performed to investigate the influence of spaceflight or simulated spaceflight environmental factors on those cognitive characteristics [14, 15]. Experiments were carried out in China's Shenzhou-9 to Shenzhou-11 space missions, and in simulated environments such as the head-down-bed-rest-experiments, the parabolic flight experiments, the isolation experiments, and the sleep restriction experiments. Those data were fitted by linear models, and the models were built in the AMSS.

By building up those characteristics models, the effects of spaceflight on human characteristics can be treated approximately in such a way that after the input of the flight state (the astronaut is on the ground, or in orbit, if in orbit, days stayed in orbit, etc.), the models predict the corresponding biomechanical and cognitive parameters, which reflect the astronaut's capabilities. Those parameters are sent to the behavioral models, and thereby may influence the human task interaction, and the performance outcome.

2.3 Behavioral Models

There are two kinds of behavioral models: biomechanical models and cognitive models. The biomechanical models are used to calculate joint forces, muscle forces, bone stress, bone mineral density in space tasks. Cognitive models are used to simulate the cognitive processes in space tasks. Through these behavioral models, we can describe human system interaction processes and get abundant behavioral data and task performance data. Those data will be further analyzed by the performance evaluation model.

Biomechanical Models

The biomechanical models and the internal relationships of these models are shown in Fig. 2. Because of weightlessness, astronaut's motion characteristics in space is different from that on ground. Especially, the phenomenon of muscular atrophy and bone remodeling during long-term spaceflight would have serious effect on the performance of astronauts. Biomechanical modeling and dynamic simulation analyses would make it possible to assess astronauts' movements and predict joint force, muscle force and bone stress. The biomechanical model includes a musculoskeletal system, a kinematics model, a kinetic model, a muscle force prediction model, and a bone stress prediction model. We constructed musculoskeletal system with 72 rigid segments which are corresponding to the anatomy structure of human. Because the shape of muscles is irregular, polygonal lines are used to replace the muscles in musculoskeletal system [16, 17]. The kinematic model takes in the joint coordinates from the musculoskeletal system, and calculates kinematic parameters (e.g. accelerated velocity, angular acceleration of body segment centroid, etc.), which become the input of the kinetic model. Based on Newton's second law, the kinetic model calculates joint forces and joint torques with input parameters and external forces (as the input parameters of biomechanical model). Similarly, muscle force prediction model calculated muscle forces based on Newton's second law. We expanded Hill model in calculating the muscle force. The difference between our model and original Hill model is the parameters of

muscle lengths and areas are variables which depend on days in orbit. With the muscle forces and joint forces, we calculate bone stress by ANSYS software using the finite element analysis method [18].

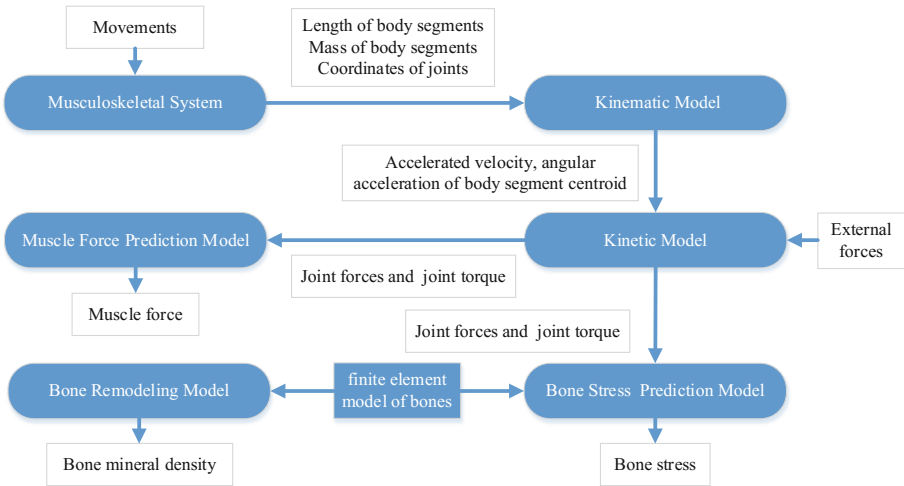


Fig. 2. The biomechanical models for simulation of space tasks

Cognitive Models

The Adaptive Control of Thought-Rational (ACT-R) has been adopted in cognitive simulations in the platform, as shown in Fig. 3. The cognitive models based on ACT-R for simulation of space tasks. Cognitive processes of specific space tasks are separated

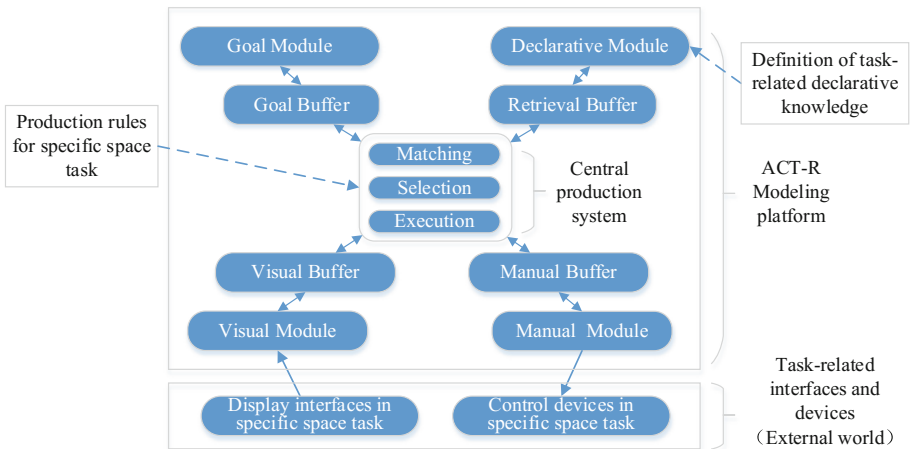


Fig. 3. The cognitive models based on ACT-R for simulation of space tasks

into different stages, such as perception, decision-making and manual operation through cognitive task analysis. Covering the stages of cognitive processes, the cognitive model includes goal module, declarative memory module, visual perception module, production system module and manual operation module. The visual perception module serves as the information input of the cognitive model while the manual operation module serves as the information output of the cognitive model. The goal module stores the dynamic goal information of the cognitive model temporarily while the declarative memory module stores the static memory of the cognitive model. The production system module is responsible for central information processing of the cognitive model and interacts with other modules through buffers.

In the cognitive model, task-related declarative memory and procedural memory, long-term memory and working memory can be represented, and relevant parameters can be provided to reflect human cognitive abilities and restrictions. Generally, different settings of these parameters will result in different performance outcome in the cognitive simulations [19]. Thus, when the models are validated and the settings of the parameters are accurate, the models have the potential to predict the performance of astronauts in specific space task with different cognitive capabilities, as well as the performance change of the same astronaut when his/her cognitive capabilities changes in the course of spaceflight, which are meaningful for the selection and training of astronauts, as well as the task arrangement in space.

2.4 Performance Evaluation Models

The output of behavioral models are usually raw process data or result data, such as the joint torque, muscle force, bone stress in the whole operation process. To get a good understanding of human performance from those data, performance evaluation models are needed. The performance evaluation models take in the process data or result data, make calculations by specific algorithms, and output more synthetic indices such as the physical workload and the comprehensive performance.

Evaluation of Physical Performance

To evaluate the physical performance, the joint workload evaluation model, fatigue evaluation model, fracture risk assessment model, and muscle strain risk assessment model were built. The relative joint torque, which is defined as the ratio of the current joint torque to the maximum joint torque of the operator, was used as an evaluation index of single joint workload. The upper limb workload evaluation model was established by synthesizing the workload of multiple joints through the method of analytic hierarchy process. In order to analyze the astronaut's fatigue in specific tasks, the joint fatigue model was established based on the movement time and energy consumption rate. The energy consumption rate is related to the joint torque and the angle of joint movement. Fatigue accumulation and fatigue recovery were considered in the model. Finally, the fracture risk assessment model and the muscle strain risk assessment model were established through the threshold comparison method [20]. The bone stress threshold was obtained through experiments and literature reports [21, 22].

Evaluation of Cognitive Performance

The indices of cognitive performance are rather task specific. In typical space tasks with relatively high cognitive demands, such as the manual rendezvous and docking (manual RVD) task, the control of space manipulator, task complete time, fuel consumption, control accuracy (usually in multiple dimensions) are basic task performance. Comprehensive models synthesizing those multiple indices were proposed and utilized in the performance evaluation in crew training and selection. Moreover, the raw process data of the cognitive models during simulation are sent to a workload evaluation model, which calculate mental workload based on the multiple resource theory [19].

2.5 Model Validation

Models in AMSS has been validated in several ways. The validation method most commonly used is to compare the outputs of the model with the test results of human participants performing the same tasks in spaceflight or simulated spaceflight environment. As some process data, such as the joint torque is hard to measure, we also make comparisons of the outputs of AMSS with some commercial software such as Visual3D, to validate models in AMSS.

Validation of the Biomechanical Models

To validate the biomechanical models, testing equipment and systems were built for biomechanical tests performed on ground, in the neutral buoyance tank, and in parabolic flights. Basic physical operations such as push, pull, lift, press, double arm rotation were performed by participants in the different environment mentioned above, and abundant data such as the body movement, operational force, support reaction force, electromyography (EMG) were collected. The body movement and operational force were input into the biomechanical models, and the models output the support reaction force and muscle activity. The support reaction force that the model output was significantly correlated to that measured in the experiment, the muscle activity that the model output was significantly correlated to the integral EMG that measured in the experiment, which proved that the models have adequate validity.

As some biomechanical data are hard to measure in experiment, we also make comparisons of the outputs of AMSS with some commercial software such as Visual3D, to validate models in AMSS. For the same set of physical tasks, biomechanical data such as the joint torque, the torque angles, the accelerated velocity, the angular speed and the angular acceleration of body segment centroid were simulated by both the biomechanical models in AMSS and the Visual3D, the data output of the two software were highly consistent, which also proved the models' validity.

Validation of the Cognitive Models

To verify the predictive abilities of the cognitive models, we chose to simulate the manual RVD task, a task commonly required of astronauts during space missions, and compare the performance outcome of the models with that of the human participants.

A research team in China Astronaut Research and Training Center has developed a simulation system of manual RVD by VC ++ language, with the modeling and simulation of the Guidance, Navigation, and Control (GNC) and system dynamics, the

docking mechanism, the instrumentation and the TV video. In the manual RVD simulator, the operator, displays, and controllers form a human-in-the-loop system [23]. The operator can observe the information displayed on the monitor and manipulates the controllers to complete the manual RVD tasks. RVD performance, such as control time and fuel consumption, is automatically recorded by the simulation system. The main task processes of manual rendezvous and docking (RVD) include: ascertain the position, attitude, and velocity of the chaser spacecraft relative to the target spacecraft, based on video images and radar information; decide on a strategy for navigating the chaser spacecraft into position; and manually control the joysticks in order to maneuver the chaser into a docking position.

Since the software of manual RVD simulator (developed by VC ++ language) and the software of cognitive models (developed by Common Lisp language) run on different platforms, it is necessary to develop the communication interface between different platforms to enable model-in-the-loop simulations. So a communication interface based on UDP and the multicast techniques was developed, which support data sending and receiving and fulfill the requirement of the real-time communication between the cognitive model and the task simulator.

Comparisons of the task performance including docking precisions, docking processes and fuel cost were made between the cognitive model and skilled operators by correlation analysis. Results showed that the correlation coefficients are above 0.8 ($p < 0.01$), which demonstrates that the cognitive models could simulate the performance of human participants well [19].

3 The Integrated Simulation Platform

The AMSS has been implemented and integrated through three layers: a user interface layer, a functional implementation layer, and a hardware layer (as shown in Fig. 4).

The user interface layer facilitates the management and scheduling of simulations and the input for setting task and model parameters. It provides the main interface for entering task-specific parameters and model parameter configurations.

The functional implementation layer enables cognitive and biomechanical simulations, human performance evaluation and analysis, task process visualization, database management and communications among multiple models. The cognitive simulation module is responsible for the simulation of human cognitive processes, while the biomechanics simulation module performs biomechanical analysis of the operator completing specific tasks in space. Task performance analyses are carried out by the performance evaluation module. The 3-D visualization module provides a geometric virtual human, task-specific images and videos necessary to visualize task execution processes [24]. The network communication module controls communication and data sharing among the multiple modules within the platform. The database module, which ties to the hardware layer, stores recorded data from simulations and performs data processing functions such as adding, modifying, deleting, querying, or browsing the data.

Finally, there is a hardware layer, which consists of a computer cluster that provides the simulation platform with high-performance calculation capabilities.

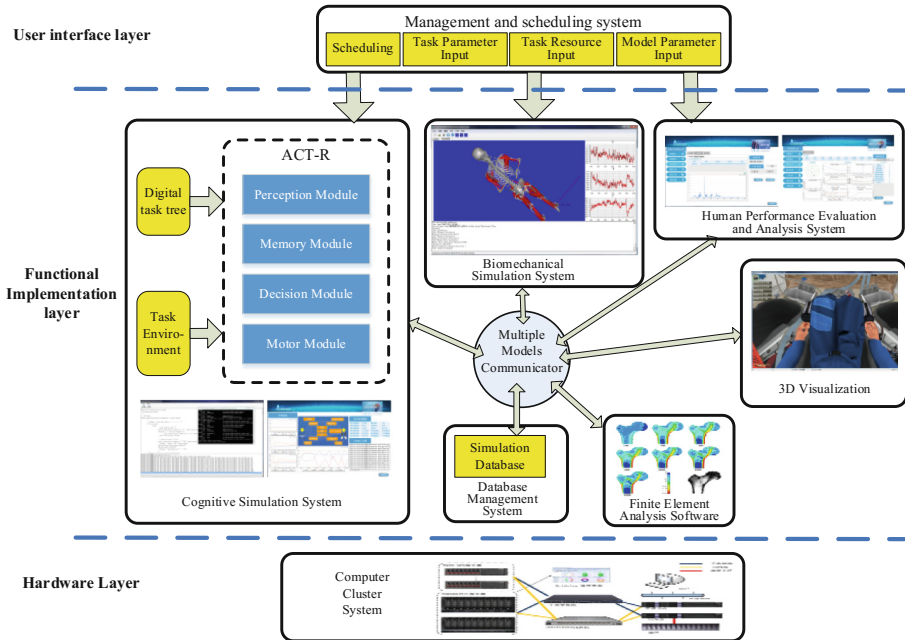


Fig. 4. The architecture of the integrated simulation platform.

The AMSS has been used to perform the quantitative evaluation of the human-machine interface designs of China's space lab and the on-going space station missions. In the ergonomic design and evaluation of the display interfaces of the manual RVD system in China's space lab missions, AMSS was employed to simulate and predict the dynamic mental workload and human performance for operators with different cognitive characteristics and for different interface schemes. Such simulations and analyses provided helpful information for the interface designs. In China's ongoing space station design, AMSS has been employed to evaluate the workload and fatigue levels of physical tasks, and proper work postures of the astronaut and the strength range requirements are derived from those simulations. AMSS will be used more widely in the ergonomic design and evaluation of China's space station missions in the coming years.

4 Conclusions

In the current paper, we introduce the Astronaut Modeling and Simulation System (AMSS), which supports modeling and simulation of astronaut's performance in specific physical and cognitive tasks during spaceflight. AMSS is the first integrated modeling and simulation platform for the human-system integration design of long-term manned space missions in China. A three-level model architecture has been proposed, which consists of the human characteristic models, the behavioral models

(cognitive and biomechanical) and the performance evaluation models. Multiple models are integrated in AMSS. The ability to visualize the virtual environment of space vehicle, the virtual astronaut, the operator's performance and task processes makes AMSS a more user-friendly platform. Models in AMSS has been preliminarily validated by experimental data. AMSS has been used to perform the quantitative evaluation of the human-machine interface designs of China's space lab and the on-going space station missions.

In the future, more data should be collected during spaceflight or simulated environment to enable better understanding of the impacts of long-term weightlessness, isolation, change of circadian rhythms and other spaceflight related factors on human characteristics. The characteristics models, the behavioral models and the performance evaluation models should be improved and enriched continuously. The software interfaces also need to be improved and updated to better satisfy the demand of the engineers in the space industry.

Acknowledgments. This work was supported by the National Basic Research Program of China (2011CB711000).

References

1. Chen, S., Wang, C., Chen, X., Jiang, G.: Study on changes of human performance capabilities in long-duration spaceflight. *Space Med. Med. Eng.* **28**(1), 1–10 (2015). (in Chinese)
2. Geuna, S., Brunelli, F., Perino, M.A.: Stressors, stress and stress consequences during long-duration manned space missions: a descriptive model. *Acta Astronaut.* **36**(6), 347–356 (1995)
3. <https://humanresearchroadmap.nasa.gov/>
4. Byrne, M.D., Pew, R.W.: A history and primer of human performance modeling. *Rev. Hum. Factors Ergon.* **5**(1), 225–263 (2009)
5. The Human Integration Design Handbook (HIDH). NASA/SP-2010-3407
6. Stambolian, D.B., Lawrence, B.A., Stelges, K.S., Ndiaye, M.O.S., Ridgwell, L.C., Mills, R.E., et al.: Human modeling for ground processing human factors engineering analysis. In: *Aerospace Conference*, pp. 1–9 (2012)
7. Delp, S.L., Anderson, F.C., Arnold, A.S., Loan, P.: Opensim: open-source software to create and analyze dynamic simulations of movement. *IEEE Trans. Bio-med. Eng.* **54**(11), 1940 (2007)
8. Anderson, J.R.: *How can the Human Mind Occur in the Physical Universe?* Oxford University Press, Oxford (2007)
9. Gore, B.F.: Man-machine integration design and analysis system (MIDAS) v5: augmentations, motivations, and directions for aeronautics applications. In: *Human Modelling in Assisted Transportation*, pp. 43–54 (2010)
10. Xiao, K., Liang, A.L., Guan, H.B., Hassanien, A.E.: Extraction and application of deformation-based feature in medical images. *Neurocomputing* **120**(10), 177–184 (2013)
11. Lei Zhou, J., Wang, D., Wang, C., Chen, S.: Remodeling models and numerical simulation of bone functional adaptation. *Chin. J. Biomed. Eng.* **33**(2), 227–232 (2014). (in Chinese)

12. Wang, J., Li, Y., Wang, F., Wang, Q., Wang, D.: Relationship between mineral density and elastic modulus of human cancellous bone. *J. Med. Biomech.* **29**(5), 465–470 (2014). (in Chinese)
13. Tian, Y.: Investigations on the key cognitive characteristics in the manually controlled rendezvous and docking task. Doctoral dissertation, China Astronaut Research and Training Center (2012). (in Chinese)
14. Chen, S., Zhou, R., Xiu, L., Chen, S., Chen, X., Tan, C.: Effects of 45-day -6° head-down bed rest on the time-based prospective memory. *Acta Astronaut.* **84**, 81–87 (2013)
15. Wang, D., Gao, W., Tian, Y., Wang, C., Ge, L., Chen, S.: Influence of gravity on speed perception characteristics of human. *Space Med. Med. Eng.* **28**(6), 408–412 (2015). (in Chinese)
16. Tang, G., Wang, C.: A muscle-path-plane method for representing muscle contraction during joint movement. *Comput. Methods Biomech. Biomed. Eng.* **13**(6), 723 (2010)
17. Garner, B.A., Pandy, M.G.: The obstacle-set method for representing muscle paths in musculoskeletal models. *Comput. Methods Biomech. Biomed. Eng.* **3**(1), 1 (2000)
18. Tang, G., Wang, D., Xiao, K., Wang, C., Chen, S.: Biomechanical modeling and dynamics simulation of an astronaut's musculoskeletal system. In: *Human Performance in Space: Advancing Astronautics Research in China (A Sponsored Supplement to Science)*, pp. 64–65 (2014)
19. Zhang, S.: Cognitive modeling research of the manual rendezvous and docking task based on ACT-R. Doctoral dissertation, China Astronaut Research and Training Center (2015). (in Chinese)
20. Li, H.: Research on virtual astronaut's physical workload evaluation based on biomechanics. Doctoral dissertation, China Astronaut Research and Training Center (2013). (in Chinese)
21. Wang, D., Shi, D., Li, X., Dong, J., Wang, C., Chen, S.: Biomechanical comparison of different pedicle screw fixations for thoracolumbar burst fractures using finite element method. *Appl. Mech. Mater.* **117–119**, 699–702 (2011)
22. Wang, S., Wang, D., Wang, F., Wang, Q., Wang, C., Chen, S.: Tensile and compressive mechanical property of human bone tissue. *Chin. J. Tissue Eng. Res.* **17**(7), 1180–1184 (2013). (in Chinese)
23. Wang, B., Jiang, G., Chao, J., Wang, X., Wang, Y., Wang, C., Lian, S.: Design and implement of manned rendezvous and docking ergonomics experimental system. *Space Med. Med. Eng.* **24**, 30–35 (2011). (in Chinese)
24. Liu, Y., Zhou, B., Zhu, X., Wang, C., Chen, S.: A visualization simulation platform of cognitive workload and performance analysis for space operations. In: *The 64th IAC International Astronautical Congress*, pp. 4577–4581 (2013)



Risk Cognition Variables and Flight Exceedance Behaviors of Airline Transport Pilots

Lei Wang^(✉), Jingyi Zhang, Hui Sun, and Yong Ren

Flight Technology College, Civil Aviation University of China, Tianjin 300300,
China
wanglei0564@hotmail.com

Abstract. In order to examine the relationship between risk cognitive variables and flight exceedance behaviors of airline transport pilots, the concepts of ‘risky pilots’ was put forward, and the existence of ‘risky pilots’ was verified based on flight exceedance events statistics. Three risk cognitive variables of pilots involving risk tolerance, hazardous attitude and risk perception were investigated through the use of a series of psychological scales. Then one-way ANOVA and Pearson Correlation Analysis were used to study the influence of risk cognitive variables on airline transport pilots’ exceedance behaviors. Results indicated that ‘risky pilots’ do exist among airline transport pilots. Additionally, a hazardous attitude has significant positive effects on pilots’ severe exceedance behaviors, which means that pilots who have high levels of a hazardous attitude would cause exceedance behaviors easily and induce higher occurrence rates of unsafe events. The conclusion of the study indicates that targeted education and training for improving risky pilots’ hazardous attitudes should be carried out for the purpose of reducing exceedance behaviors in flight, which would then make flight safer.

Keywords: Risk cognition · Exceedance behaviors · Risky pilots
Flight data · Safety

1 Introduction

According to the accident proneness theory [1], there are great individual differences among human beings. The viewpoint that some people who possess accident proneness are essentially more prone to accidents compared to others due to their physiological and psychological differences has been widely recognized in the automotive industry. The research on the correlation between driving behaviors, accidents and personality characteristics of accident-prone drivers has also been widely conducted [2, 3].

Since the 1990s, aeronautical psychologists have begun to focus on the impact of risk factors in personality traits of pilots on flight safety. Hunter [4] called the factor risk tolerance, which refers to the quantity and extent of willingness to accept the risk in order to achieve a particular goal. Pauley et al. [5] held that for the purpose of accomplishing specific tasks, or achieving specific goals, pilots may need to tolerate or accept high-level risk to execute flight plans while flying. Related studies have shown

that there is a close interconnection between risk tolerance and the safety performance of pilots [6–8], and a negative correlation between the risk tolerance level and the safety behaviors of pilots [4]. Berlin et al. [9] initially offered that hazardous thought patterns and the influence of pilots' hazardous attitudes on driving safety behaviors have been deliberately examined by researchers. The Federal Aviation Administration (FAA) found that hazardous attitudes, such as manliness, anti-authoritarian attitudes, impulsiveness, compliance and tenacity could affect the level of decision-making of pilots [10]. Ji et al. [11] noticed that hazardous attitudes promote the negative effects of risk tolerance on flight safety. Meanwhile, risk perception is considered as a type of situational risk awareness, individuals' subjective cognition and evaluation of potential hazards in their external environment, and corresponding preparatory behavior [12], also attracted concern from psychologists. Hunter [4] found that risk perception is negatively correlated with risk tolerance among pilots. Ji et al. [11] posited that risk perception plays a regulatory role in the process of risk tolerance, affecting pilots' driving safety behaviors. Ji et al. [11] also suggested that a high level of risk perception will weaken the negative effect of risk tolerance on driving safety behaviors. Additionally, the study showed that such risk perception will also regulate the effect of hazardous attitudes on the impact of risk tolerance on pilots' driving safety behaviors.

An exceedance event is an unsafe event in which the monitoring parameters of any quick access recorder exceed the standard of Flight Operations Quality Assurance (FOQA), specifically aiming at the "collection and analysis system of flight data in daily flight" [13, 14], and is independently reported by the Flight Operations Quality Assurance software. Exceedance event risk management based on Quick Access Recorder (QAR) data is the core content of the FOQA of airlines. Similar to the idea of big data, the flight data of QAR could be used to analyze and evaluate the operant level and exceedance behavior of pilots. Wang et al. [15–19] carried out research on the evaluation of flight operation risk by using QAR data, especially the in-depth study of flaring operational characteristics of long landing and hard landing events, which specifically analyze the impact of flaring operation on landing performance. Overall, related research on exceedance events using QAR data that pertains to the detection, diagnosis, and prediction of exceedance events but lack of studies on the relationship between risk-related personality traits and exceedance behaviors.

Therefore, statistics and analysis of the exceedance events of airline transport pilots will be conducted, to verify whether the minority, who are more prone to exceedance events compared to others, exist among airline transport pilots as motorists. In addition, differences of QAR exceedance events of airline transport pilots, at different levels of risk cognition, will be analyzed. The results can provide a theoretical reference for the Flight Operations Quality Assurance work and selection of airline transport pilots.

2 Methods

2.1 Exceedance Events Statistics

In order to verify the existence of "risky pilots"—there are some pilots who are more prone to exceedance events compared to others—a study of the exceedance cases

(including mild exceedance and severe exceedance) of airline transport pilots within the past one year were gathered and analyzed, and then contrasted in detail. The findings of the types of specific exceedance events and the occurrence rate of each exceedance event are shown in Table 1.

Table 1. Statistics of exceedance events

| Exceedance event | Jan. | Feb. | Mar. | Apr. | May. | Jun. | Jul. | Aug. | Sept. | Oct. | Nov. | Dec. |
|---------------------------------------|------|------|------|------|------|------|------|------|-------|------|------|------|
| Straight taxiing overspeed | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Cornering taxiing overspeed | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| Twin-engined torque deviation | 1 | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 5 | 4 |
| Exceeding tire limit speed | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| Connecting autopilot early | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| Squash in initially climbing | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| High descent rate (2000–1000ft) | 3 | 1 | 4 | 2 | 3 | 1 | 0 | 3 | 1 | 0 | 0 | 0 |
| High descent rate (1000–500ft) | 12 | 5 | 21 | 11 | 18 | 4 | 3 | 17 | 10 | 11 | 7 | 12 |
| High descent rate (500–50ft) | 10 | 5 | 18 | 8 | 17 | 2 | 0 | 8 | 7 | 6 | 5 | 5 |
| Low touch-down speed | 66 | 98 | 65 | 52 | 17 | 31 | 24 | 19 | 28 | 44 | 19 | 28 |
| Small touch-down pitch | 30 | 39 | 32 | 47 | 51 | 14 | 17 | 29 | 21 | 29 | 50 | 80 |
| Large flare landing bank | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| Large vertical load during landing | 3 | 6 | 5 | 4 | 4 | 5 | 7 | 2 | 4 | 4 | 2 | 3 |
| Long touchdown distance from 50ft | 19 | 21 | 15 | 30 | 21 | 22 | 31 | 39 | 27 | 21 | 33 | 33 |
| Large vertical load during air | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| GLIDESLOPE warning | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| SINK RATE warning | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

On the basis of obtaining the statistical information of the airline's exceedance events within the past year, an correspondence statistic for each pilots' exceedance times were conducted further, and the pilots who undertook the task of operation in each flight were confirmed through the airline's flight management software. If the pilot completed a take-off or a landing operation as the pilot in operation, it was recorded accordingly. However, the take-off or landing operations of the non-main operators were not included in the exceedance times.

In order to analyze the exceedance cases of airline transport pilots with different risk cognition, occurrence rate of exceedance events—exceedance rate being the ratio of the cumulative times of exceedance events over a year to the overall take-off and landing—was put forward to evaluate the risk behavior tendency of airline transport pilots as an index. For instance, the findings discovered that the higher the exceedance rate was evaluated, the higher the tendency of risky behavior. After the statistics of each pilots' exceedance times were recorded within the past year, the number of flights during the year were analyzed.

2.2 Measurement of Risk Cognitive Variables

In order to measure the degree of pilots' risk tolerance, risk perception and hazardous attitude shown in flight, the total scale of airline transport pilots' risk cognition were established by consultation and translation of the risk tolerance scale of pilots, risk perception scale of pilots, and hazardous attitude scale of pilots, compiled by Hunter [4] and Ji et al. [11].

2.2.1 Scale Structure

- (1) Risk Tolerance. The airline transport pilots' risk tolerance scale is made up of 17 kinds of flight situations, which were established according to the study of Hunter [4] by Ji et al. [11]. The total scale of risk tolerance consists of 3 parts: risk tolerance for aircraft system failure (3 topics), risk tolerance for crew operation (7 topics), and risk tolerance for flight environment (7 topics). The risk tolerance score is measured in 5 grades: 1. pilots who said they are extraordinarily willing to accept, or to agree with, the flight situation given in the scale; 2. pilots who said they are willing to accept, or to agree with, the flight situation given in the scale; 3. pilots who said they are generally willing to accept, or to agree with, the flight situation given in the scale; 4. pilots who said they are reluctant to accept, or to agree with, the flight situation given in the scale; and 5. pilots who said they are very reluctant to accept, or to agree with, the flight situation given in the scale. The score of the scale is the average of all the topics; hence, the higher the score, the higher the level of risk tolerance.

According to Ji et al. [11], the reliability and validity of the test results of the scale reveal that the Cronbach's α coefficient on the scale is 0.81, whereas the test results of the scale translated to the Cronbach's α coefficient is 0.94.

- (2) Risk Perception. The risk perception scale of airline transport pilots consists of 26 kinds of flight situations or flight events. It has been widely applied in pilots' research since it was compiled by Hunter [4]. The total risk perception scale

consists of 5 parts: general flight risk (5 topics), high flight risk (7 topics), flight altitude risk (7 topics), automobile driving risk (3 topics) and daily life risk (4 topics). The grades of risk situations, or risk events, listed in the scale are 0–100. The score of the scale is the average of all the topics. Hence, the higher the score, the higher is the level of risk perception.

According to Ji et al. [11], the test results of the reliability and validity of the scale exhibit that the Cronbach’s α coefficient of scale is 0.89, whereas the test results of the scale translated to that of the Cronbach’s α coefficient is 0.910.

- (3) Hazardous Attitude. The airline transport pilots’ hazardous attitude scale consists of 24 kinds of behaviors that are closely related to modern airline activities. The study was conducted in China by Ji et al. [20]. The hazardous attitude scale consists of 6 parts: confidence (6 topics), impulses (5 topics), manliness (3 topics), anxiety (4 topics), obedience (3 topics) and risk awareness (3 topics). The scale is a Likert-type five point scale. 1. pilots who said they are extraordinarily willing to accept, or to agree with, the flight situation given in the scale; 2. pilots who said they are willing to accept, or to agree with, the flight situation given in the scale; 3. pilots who said they are generally willing to accept, or to agree with, the flight situation given in the scale; 4. pilots who said they are reluctant to accept, or to agree with, the flight situation given in the scale; and 5. pilots who said they are very reluctant to accept, or to agree with, the flight situation given in the scale. The score of the scale is the average of all the topics. Hence, the higher the score, the higher is the level of the hazardous attitude.

According to Ji et al. [11], the reliability and validity of test results of the scale exhibit that the Cronbach’s α coefficient of scale is 0.89, whereas the test results of the scale translated to that of the Cronbach’s α coefficient is 0.912.

2.2.2 Scale Implementation

After the distribution of the questionnaires, 66 completed questionnaires were collected from the airline. The basic statistical data of the subjects are shown in Table 2.

Table 2. Basic statistical data of subjects

| | Hierarchy | Number | Proportion |
|--------|-----------|--------|------------|
| Age | 20–25 | 4 | 6.06% |
| | 26–30 | 17 | 25.76% |
| | 31–35 | 19 | 28.79% |
| | 36–40 | 7 | 10.61% |
| | 41–45 | 8 | 12.12% |
| | 46–50 | 6 | 9.09% |
| | Over 50 | 5 | 7.58% |
| Gender | Male | 66 | 100.00% |
| | Female | 0 | 0.00% |

(continued)

Table 2. (continued)

| | Hierarchy | Number | Proportion |
|--------------|-----------------|--------|------------|
| Position | Instructor | 4 | 6.06% |
| | Captain | 8 | 12.12% |
| | First officer | 48 | 72.73% |
| | Trainee | 6 | 9.09% |
| Flight hours | Less than 1000 | 26 | 39.39% |
| | 1000–2000 | 16 | 24.24% |
| | 2000–4000 | 12 | 18.18% |
| | 4000–6000 | 4 | 6.06% |
| | 6000–8000 | 1 | 1.52% |
| | 8000–10000 | 3 | 4.55% |
| | 10000–12000 | 1 | 1.52% |
| | More than 12000 | 3 | 4.55% |

Of the 66 completed questionnaires, 51 agreed with using the individual's real name to participate in the follow-up research and analysis. Also, the participants polled agreed to have their exceedance records taken from the Flight Operations Quality Assurance system of their airline. Sample selection was based on the following conditions: (1) at least one year of service time; (2) more than 100 h of flight hours within the past year; (3) no occurrences of accidents or accidental signs occurring as the main operator; (4) more than 50 (including 50) sorties per year in total. Five out of 51 questionnaires did not meet the screening criteria. The scores of each sample were calculated separately, and then matched one-to-one with their exceedance rate. Finally, 46 effective samples were considered.

2.2.3 Sample Analysis

After obtaining 46 effective samples, with the aim of examining the relationship between risk cognitive variables and the exceedance situations of airline transport pilots, one-way ANOVAs and Pearson Correlation Analyses were conducted through SPSS data analysis software.

In relation to studying the difference shown by pilots' risk cognitive variables at different exceedance rate levels, the mild and severe exceedance rates of the 46 samples were divided into three groups: high ($M + SD$), medium (M), low ($M - SD$). The boxplot indicates that outliers exist in the samples, but there are no extreme outliers. In order to ensure the integrity and authenticity of the samples, the outlier cannot be excluded. Based on the Shapiro-Wilk test, each group of data obeys a Gaussian distribution ($P > 0.05$). Through Levene's test of homogeneity of variances, each group was shown to be consistent ($P > 0.05$).

In order to examine the correlation coefficient between the hazardous attitude and the severe exceedance rate further, Pearson Correlation Analysis was carried out to analyze the relationship between risk cognitive variables and exceedance rates. By scattergram, there is no obvious outlier in the data. However, in reference to the Shapiro-Wilk test, all the continuous variables conform to a Gaussian distribution ($P > 0.05$), and there is no obvious skewness in the 46 samples.

3 Results

3.1 Verification of Risky Pilots

The population composition proportion and exceedance times proportion were compared, as is exhibited with their corresponding outcomes in Fig. 1.

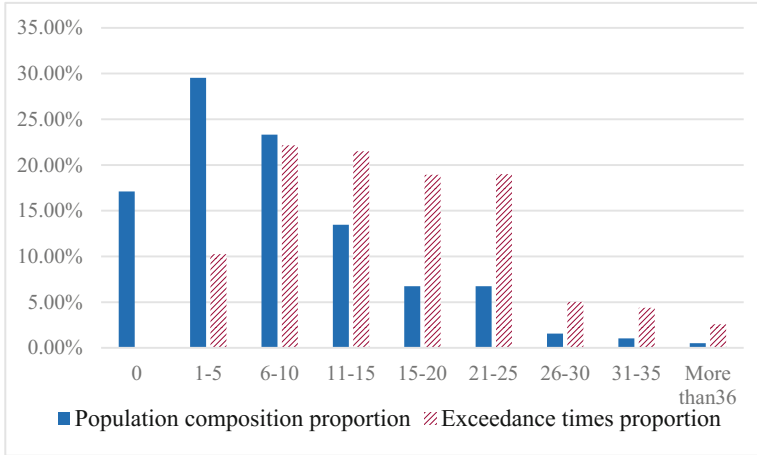


Fig. 1. Comparison of population composition proportion and exceedance times proportion

As shown in Fig. 1:

- (1) There are 193 pilots in the airlines. Thirty-three pilots did not take part in exceedance events within the past year, which accounted for 17.10% of the total; whereas, 160 pilots did take part in exceedance events, accounting for 82.90%.
- (2) Pilots whose exceedance events were less than 10 times (including 10 times) within the past year accounted for 69.95% of the total. The 69.9% pilots' population composition proportion was higher than the exceedance times proportion. Pilots whose exceedance events were more than 10 times within the past year accounted for 30.05% of the total. The 30.05% pilots' population composition proportion was lower than the exceedance times proportion. The more intuitive results can be shown in Figs. 2 and 3.

By percentage analysis, a preliminary comparative result of exceedance frequency within the past year could be produced. Thus, a small number of pilots in the airline accounted for a large proportion of the total exceedance events, and their exceedance rates are significantly higher than other pilots', which proves that a specified minority of pilots often take part in exceedance events. That is to say, these pilots are more prone to exceedance events than others.

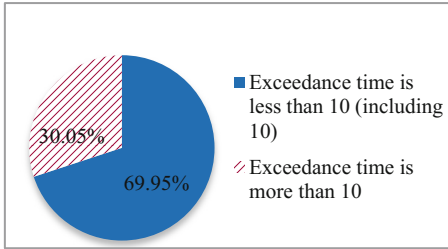


Fig. 2. Population composition proportion

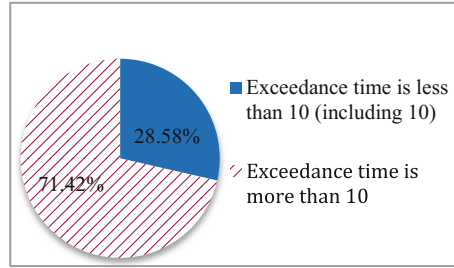


Fig. 3. Exceedance times proportion

3.2 Correlation Between Risk Cognitive Variables and Exceedance Behaviors

The one-way ANOVA exhibited that there were significant differences in hazardous attitudes among the groups with different levels of severe exceedance rates, and the difference was statistically significant; $F(2, 43) = 3.497, P < 0.05$. According to the Tukey-Kramer multiple comparison post test results, the following could be seen: the hazardous attitude score of the “high severe exceedance rate” group was 0.44 higher than the “low severe exceedance rate” group (95% CI: 0.01–0.87), and the difference was statistically significant ($P < 0.05$). The specific results are as shown in Table 3 and Fig. 4.

Table 3. Results of ANOVA

| Variables | Sum of squares | df | Mean square | F | Sig. | |
|--------------------|----------------|-----------|-------------|---------|-------|--------|
| Risk tolerance | Between groups | 0.511 | 2 | 0.256 | 0.846 | 0.436 |
| | Within groups | 12.991 | 43 | 0.302 | / | / |
| | Total | 13.502 | 45 | / | / | / |
| Risk perception | Between groups | 955.747 | 2 | 477.874 | 2.086 | 0.137 |
| | Within groups | 9852.277 | 43 | 229.123 | / | / |
| | Total | 10808.024 | 45 | / | / | / |
| Hazardous attitude | Between groups | 0.937 | 2 | 0.468 | 3.497 | 0.039* |
| | Within groups | 5.760 | 43 | 0.134 | / | / |
| | Total | 6.697 | 45 | / | / | / |

The results of Pearson Correlation Analysis showed that there was a moderate positive linear correlation between hazardous attitudes and severe exceedance rates, $R(44) = 0.400, P < 0.05$. The specific results are shown in Table 4.

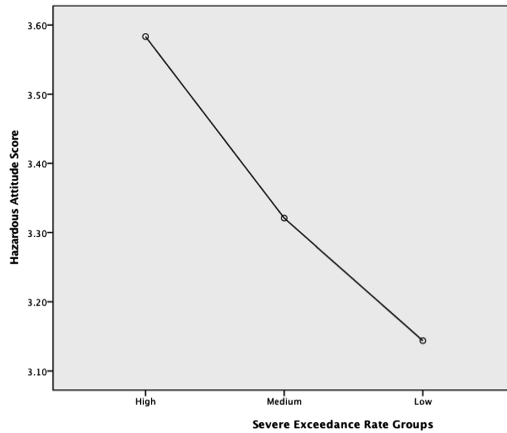


Fig. 4. The differences of hazardous attitudes at different levels of severe exceedance rate

Table 4. Mean, standard deviation and correlation coefficient of variables

| Variables | <i>M</i> | <i>SD</i> | 1 | 2 | 3 | 4 | 5 |
|------------------------|----------|-----------|--------|--------|---------|-------|---|
| Risk tolerance | 1.9910 | 0.5478 | — | | | | |
| Risk perception | 48.1246 | 15.4977 | -0.142 | — | | | |
| Hazardous attitude | 3.2301 | 0.3858 | 0.171 | 0.198 | — | | |
| Mild exceedance rate | 0.0660 | 0.0640 | -0.187 | -0.003 | 0.212 | — | |
| Severe exceedance rate | 0.0031 | 0.0064 | 0.087 | 0.212 | 0.400** | 0.455 | — |

4 Discussion

4.1 Existence of Risky Pilots

Accident proneness theory posits that some people are inherently more prone to accidents than others in the same environment [1]. This study shows that there are a small percentage of people whose exceedance times in the last three years number far more than others in the airline, after the airline's frequency of exceedance times in the last three years was compared initially using the percentage method. Statistics and tests analyzed regarding the distribution of exceedance times and exceedance rates of pilots driving two different types of aircraft show that the distribution of exceedance times of pilots exhibited an obvious "right deviation" configuration, whereas the distribution of exceedance rates remained the same. This partly reflects the fact that the occurrence of exceedance events is biased towards a specific group of people. The study shows that there is a significant correlation between exceedance rates of pilots in different periods, which means the exceedance rates of pilots with a high exceedance rate in the previous stage will be higher than the exceedance rate in the later stage. The exceedance rate of pilots is a relatively stable indicator. These results show that the occurrence of

exceedance events is not completely random, like traffic accidents in the driving area, but rather is attributable to a particular group who, when compared to other pilots, were more likely to exceed the exceedance rates. Hence, there are “risky pilots” in the pilot population.

4.2 The Positive Effect of Hazardous Attitudes on Severe Exceedance Behaviors

Hazardous attitudes are influenced by personal needs and the external environment. It is undeniable that attitudes are unstable and subjective but will affect the behaviors of individuals with respect to the population, environment and various situations, whether or not they can be manifested as a kind of perception that has been known for a considerable amount of time (Wilkening 1973) [21]. The results show that the level of hazardous attitudes with a high severe exceedance rate is significantly higher than that with a low severe exceedance rate ($F(2, 43) = 3.497, p < 0.05, \eta^2 = 0.468$). Whereas the scores of hazardous attitudes decrease in the order of a high severe exceedance rate group, a middle severe exceedance rate group, and a low severe exceedance rate group, as shown in Fig. 2, Tukey’s test results show that the level of hazardous attitudes decreases by 0.26 (95% CI: $-0.22414-0.7491$), with no statistical significant difference between the high severe exceedance rate group and the middle severe exceedance rate group. Hence, the level of hazardous attitudes decreases by 0.18 (95% CI: $-0.14610-0.5001$), with no statistically significant difference between the middle severe exceedance rate group and the low severe exceedance rate group. As the level of hazardous attitudes decreases by 0.44 (95% CI: $-0.22414-0.7491$), it does so with the statistically significant difference between the high severe exceedance rate group and the low severe exceedance rate group. Hence, this conclusion indicates that the level of hazardous attitudes of pilots with high severe exceedance rates is significantly more than that of pilots with low severe exceedance rates. This is consistent with Hunter’s study [23], which exhibits that pilots with high levels of hazardous attitudes will have even more risk events. The research of Pauley et al. [5] also finds that hazardous attitudes can effectively predict the adventure tendency and dangerous aviation activities of pilots.

The results of Pearson Correlation Analysis show that there is a significant positive correlation between hazardous attitudes and severe exceedance rates, which are consistent with the results of one-way ANOVA. Hence, hazardous attitudes can be used as a valid indicator for predicting such things as the pilots’ personality traits, exceedance situations, and risk events, to effectively evaluate the flight performance of a pilot and for use as screening criteria in the selection of flight students.

4.3 The Impact of Risk Tolerance and Risk Perception on Exceedance Behaviors

However, the risk tolerance of pilots doesn’t show statistically significant correlation with mild exceedance rates and severe exceedance rates, neither does risk perception. It is similar to the findings of Hunter [22] that no significant relationship between risk tolerance and pilot risk driving behavior. Also, high levels of risk perception will

significantly affect the impact of risk tolerance on driving safety behavior. Additionally, as shown by Hunter and Stewart [23], hazardous attitudes and risk tolerance can be used as independent variables to explain an individual's risk-behavior predisposition. However, with that stated, hazardous attitudes are more useful for explaining risky driving than risk tolerance. Studies by Ji et al. [11] also show that risk tolerance has a significant positive effect on hazardous attitudes. Therefore, risk tolerance plays a direct role in hazardous attitudes and thus affects pilots' risk behaviors through the adjustment of risk perception. Risk perception also plays a role in the influence of hazardous attitudes on risky driving behavior. Since hazardous attitudes affect the pilots' severe exceedance rates, it is necessary to reduce the pilots' hazardous attitudes through training and to then reduce their severe exceedance rate. Concomitantly, such a reduction to exceedance rate will improve the safe level of a pilot's operation, including such aspects as flight system knowledge, situational awareness, risk decision-making and other non-technical skills. This result supports the findings of Berlin [9], in which training in hazardous attitudes can significantly improve the decision-making of pilots in the short term. (Federal Aviation Administration 1991; Buch and Diehl 1984; Diehl 1991) [10, 24, 25]. The hazardous attitudes of pilots include the six dimensions of confidence, impulses, manliness, anxiety, obedience and risk awareness. Exceedance rates reflect the risk tendencies of the pilots only from the event level itself, and the follow-up study would further examine the impact of each dimension of risk attitude on the risk behaviors of pilots themselves. Concomitantly, it is also more valuable to extend the research object to the relationship between the QAR data corresponding to the risk-related personality traits and the pilots' operational behaviors.

5 Conclusions

This study examined the relationship between risk cognitive variables and flight exceedance behaviors of airline transport pilots. Three risk cognitive variables of pilots involving risk tolerance, hazardous attitude and risk perception were investigated through the use of a psychological scale.

There is a population group of "accidental drivers" in the pilot community, which we define as "risky pilots." Although the "risky pilots" account for only a small percentage of the total pilot population, they trigger typical exceedance warnings at a higher frequency than their counterparts, meaning they are more prone to exceedance events than other pilots.

The pilots with high severe exceedance rates have generally higher hazardous attitudes scores than the pilots with a low severe exceedance rate. However, there is a moderate positive correlation found between the hazardous attitudes and severe exceedance rates of pilots.

It is suggested that targeted education and training for improving risky pilots' hazardous attitudes is a good way for reducing exceedance behaviors in flight.

Acknowledgments. We appreciate the support of this work from the National Natural Science Foundation of China (Grant No. U1733117) and the National Key Research and Development Program of China (Grant No. 2016YFB0502405).

References

1. Greenwood, M.: Accident proneness. *Biometrika* **37**(1/2), 24–29 (1950)
2. Li, Z., Sun, J.T., Xu, K.H., Chen, Y.X., Zhao, S.M.: Predicting accident proneness of pilot with Eysenck personality questionnaire. *Chin. J. Aerosp. Med.* **10**(4), 234–236 (2007)
3. Yan, H.: Building model for relationship between road traffic accident and drivers' psychological quality. *China Saf. Sci. J.* **26**(2), 13–17 (2016)
4. Hunter, D.R.: Risk perception and risk tolerance in aircraft pilots (Report No. DOT/FAA/AM-02/17). Federal Aviation Administration Office of Aviation Medicine, Washington DC (2002)
5. Pauley, K., O'Hare, D., Wiggins, M.: Risk tolerance and pilot involvement in hazardous events and flight into adverse weather. *J. Saf. Res.* **39**(4), 403–411 (2008)
6. O'Hare, D.: Pilots' perception of risks and hazards in general aviation. *Aviat. Space Environ. Med.* **61**(7), 599–603 (1990)
7. Platenius, P.H., Wilde, G.J.: Personal characteristics related to accident histories of Canadian pilots. *Aviat. Space Environ. Med.* **60**(1), 42–45 (1989)
8. Wiggins, M., Connan, N., Morris, C.: Weather-related decision making and self-perception amongst pilots. In: Haywood, B.J., Lowe, A.R. (eds.) *Applied Aviation Psychology: Achievement, Change and Challenge, Proceedings of 3rd Australian Psychology Symposium*, pp. 193–200. Avebury Ashgate Publishing Ltd., Aldershot (1996)
9. Berlin, J.I., Gruber, E.V., Holmes, C.W., Jensen, P.K., Lau, J.R., Mills, J.W.: Pilot judgment training and evaluation (Report No. DOT/FAA/CT-81/56-I), vol. 1. Federal Aviation Administration, Washington, DC (1982)
10. Federal Aviation Administration: Aeronautical decision making. Advisory Circular: 60–22 FAA, Washington, DC (1991)
11. Ji, M., You, X.Q., Lan, J.J., Yang, S.Y.: 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)
12. Tränkle, U., Gelau, C., Metker, T.: Risk perception and age-specific accidents of young drivers. *Accid. Anal. Prev.* **22**(2), 119–125 (1990)
13. Civil Aviation Administration of China: Implementation and management of flight operation quality assurance. Advisory Circular: 121/135-FS-2012-45. CAAC, Beijing (2012)
14. Yang, Y.X.: Research on flight operation risk based on quick access recorder data. Unpublished Master's dissertation, Civil Aviation University of China (2016)
15. Wang, L., Sun, R.S., Wu, C.X., Cui, Z.X., Lu, Z.: A flight QAR data based model for hard landing risk quantitative evaluation. *China Saf. Sci. J.* **24**(2), 88–92 (2014)
16. Wang, L., Wu, C.X., Sun, R.S.: An analysis of flight Quick Access Recorder (QAR) data and its applications in preventing landing incidents. *Reliabil. Eng. Syst. Saf.* **127**, 86–96 (2014)
17. Wang, L., Wu, C., Sun, R., Cui, Z.: An analysis of hard landing incidents based on flight QAR data. In: Harris, D. (ed.) *EPCE 2014. LNCS (LNAI)*, vol. 8532, pp. 398–406. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07515-0_40
18. Wang, L., Wu, C., Sun, R.: Pilot operating characteristics analysis of long landing based on flight QAR data. In: Harris, D. (ed.) *EPCE 2013. LNCS (LNAI)*, vol. 8020, pp. 157–166. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39354-9_18
19. Wang, L., Ren, Y., Wu, C.X.: Effects of flare operation on landing safety: a study based on ANOVA of real flight data. *Saf. Sci.* **102**, 14–25 (2018)
20. Ji, M., Liu, Z., Yang, S.Y., Bao, X.H., You, X.Q.: A study on the relationship between hazardous attitudes and safe operation behaviors among airline pilots in China. *J. Psychol. Sci.* **35**(1), 202–207 (2012)

21. Wilkening, H.E.: *The Psychology Almanac*. Brooks/Cole, Monterey (1973)
22. Hunter, D.R.: Measurement of hazardous attitudes among pilots. *Int. J. Aviat. Psychol.* **15** (1), 23–43 (2005)
23. Hunter, D.R., Stewart, J.E.: Locus of control, risk orientation, and decision making among U.S. army aviators. Technical report No. 1260, Army Research Institute for the Behavioral and Social Sciences (DTIC No. ADA452056), Arlington (2009)
24. Buch, G., Diehl, A.: An investigation of the effectiveness of pilot judgment training. *Hum. Fact.: J. Hum. Fact. Ergon. Soc.* **26**(5), 557–564 (1984)
25. Diehl, A.E.: The effectiveness of training programs for preventing aircrew ‘error’. In: Jensen, R.S. (eds.) *Proceedings of 6th International Symposium on Aviation Psychology*, pp. 640–655. Ohio State University, Columbus (1991)

Author Index

- Agnarsson, Arnar 615
Albayram, Yusuf 369
Aoki, Hirofumi 445
Atkinson, Robert 403
- Bach, Cedric 593
Bai, Peng 100
Beyerer, Jürgen 419
Böhm, Patricia 429
Boy, Guy A. 701
Braithwaite, Graham 137, 583
Brand, Yannick 3
Brandt, Summer L. 215
Buck, Ross 369
- Calvet, Guillaume 593
Camachon, Cyril 285
Castillo-Medina, Gustavo 306
Cheema, Baljeet Singh 265
Chen, Allen C. 389
Chen, Congzhe 533
Chen, Jintao 605
Chen, Shanguang 713
Chen, Xingyu 490
Chen, Xinyi 55
Coman, Emil 369
Coronado, Braulio 171
Coyne, Joseph T. 296
- Debue, Nicolas 403
Ding, Lin 241
Donath, Diana 42
Dudzik, Kate 501
- E, Xiaotian 147
Eibl, Maximilian 429
- Feng, Chuanyan 115, 241
Feng, Wenjuan 127
Ferrari, Vincent 285
Flynn, Shannon R. 296
Friedrich, Maik 558
Fu, Shan 55, 80, 317, 336
- Gagnon, Jean-François 285
Galarza-Del-Angel, Javier 306
Gale, Jack 545
Galindo-Aldana, Gilberto 306
Gao, Qin 19
Gavish, Nirit 471
Gustafson, Eric 171
Gutzwiller, Robert S. 42
- Hagl, Maria 558
Hammer, Jan-Hendrik 419
Harris, Don 572
He, Xueli 241
Hild, Jutta 419
Hillenius, Steve 545
Hinterleitner, Bastian 429
Horesh, Eran 471
Horn, Andreas 137
Hou, Tingting 317, 336
Huang, Shoupeng 713
- Isemann, Daniel 429
- Jakobi, Jörn 558
Jensen, Theodore 369
Jin, Xiaoping 90, 250
Johnston, Joan 230
- Kaiyuan, Song 352
Kaptein, Frank 204
Karasinski, John 545
Karimi, Fraydon 501
Kearney, Peter 583
Khan, Mohammad Maifi Hasan 369
Kim, Jung Hyup 32
Klaus, Edmund 419
Kopf, Maëlle 285
Koteskey, Robert 215
Krisher, Hagit 471
- Lachter, Joel 215
Lange, Douglas S. 42, 171
Large, Anne-Claire 593

- Ledesma-Amaya, Israel 306
 Li, Ding 352
 Li, Jingqiang 158
 Li, Kang 69, 158
 Li, Wen-Chin 137, 583
 Li, Yazhe 147
 Li, Yongnan 330
 Lin, Runing 90
 Liu, Bao 330
 Liu, Bisong 330
 Liu, Sha 605
 Liu, Shuang 115, 241
 Liu, Xia 330
 Liu, Yuqing 713
 Liu, Zhen 69
 Lu, Yanyu 55, 80, 317, 336
 Lucero, Crisrael 171
 Lugo, Ricardo G. 181
 Lux, Benedikt 429
 Lv, Chunhui 147
- Ma, Xuechao 250
 Maehigashi, Akihiro 445
 Martin, Manuel 419
 McCarthy, Pete 615
 McDermott Ealding, Claire 637
 Meza-Kubo, Victoria 306
 Miao, Chongchong 241
 Milham, Laura 230
 Miller, Christopher A. 191
 Miwa, Kazuhisa 445
 Morán, Alberto L. 306
 Moran, Sabrina 389, 457
- Nagy, Nathan 501
 Nare, Matthew 457
 Neerinx, Mark A. 204
- Padilla-López, Alfredo 306
 Papenfuss, Anne 558
 Pei, Yeqing 250
 Peinsipp-Byma, Elisabeth 419
 Phillips, Henry 230
 Pradhan, Anish 403
- Qian, Chen 317, 336
 Qiao, Han 147
 Qu, Xingda 490
 Quartuccio, Jacob S. 296
- Rambau, Tim 558
 Ren, Yong 725
 Riddle, Dawn 230
 Ross, William A. 230
- Samanta, Debasis 265
 Samima, Shabnam 265
 Sarma, Monalisa 265
 Scherer-Negenborn, Norbert 558
 Schmidl, Daniel 429
 Schmidt, Markus 558
 Schmitt, Fabian 649
 Schulte, Axel 3, 42, 649
 Sevillian, Dujuan 664
 Shamilov, Elias 471
 Shi, Xue 90
 Shively, Robert J. 215
 Sibley, Ciara 296
 Skovron, Ezekiel 457
 Song, Bingbing 55, 80
 Song, Zhenghe 250
 Stedmon, Alex 637
 Sun, Hui 725
 Sun, Ruishan 69, 127, 158, 686
 Sun, Xianghong 147
 Sun, Yuting 389, 480
 Sütterlin, Stefan 181
 Suzuki, Tatsuya 445
- Tan, Wei 701
 Tao, Da 490
 Teng, Xiaobi 55, 80
 Tian, Yu 713
 Townsend, Lisa 230
- van de Leemput, Cécile 403
 van der Waa, Jasper 204
 van Diggelen, Jurriaan 204
 Voit, Michael 419
 Vu, Kim-Phuong L. 389, 457, 480
- Wang, Chunhui 713
 Wang, Dongmei 713
 Wang, Haiyan 533
 Wang, Lei 725
 Wang, Qun 90
 Wang, Tieyan 490

- Wang, Xinglong 100
Wang, Yuhui 241
Wang, Zhen 55, 80
Wanyan, Xiaoru 115, 241
Ward, Lawrence 501
West, Robert 501
Wu, Man 19
Wu, Xu 115, 241
- Xiao, Youyu 330
Xie, Fang 90
Xiong, Lin 147
Xue, Chengqi 533
- Yandong, Wang 352
Ye, Hai 55, 80
Yi, Ding 352
Yuan, Juan 490
Yuan, Zhibo 686
- Zhang, Guanchao 686
Zhang, Jingyi 725
Zhang, Jingyu 147
Zhang, Kai 127
Zhang, Xingjian 100
Zhang, Youxue 605
Zhang, Yu-Ting 158
Zhang, Yuting 69
Zhao, Guozhen 520
Zhao, Yifei 100
Zheng, Bowen 250
Zheng, Yan 520
Zhongji, Zhang 352
Zhou, Xiaozhou 533
Zhou, Yi 55, 80
Zhu, Bin 19
Zhu, Zhaowei 352
Zhuang, Damin 115
Zou, Xiangying 147