Don Harris (Ed.)

LNAI 11571

Engineering Psychology and Cognitive Ergonomics

16th International Conference, EPCE 2019
Held as Part of the 21st HCI International Conference, HCII 2019
Orlando, FL, USA, July 26–31, 2019, Proceedings

Springer

Lecture Notes in Artificial Intelligence 11571

Subseries of Lecture Notes in Computer Science

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

Don Harris (Ed.)

Engineering Psychology and Cognitive Ergonomics

16th International Conference, EPCE 2019
Held as Part of the 21st HCI International Conference, HCII 2019
Orlando, FL, USA, July 26–31, 2019
Proceedings

Springer

Editor
Don Harris
Coventry University
Coventry, UK

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-030-22506-3 ISBN 978-3-030-22507-0 (eBook)
https://doi.org/10.1007/978-3-030-22507-0

LNCS Sublibrary: SL7 – Artificial Intelligence

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

Foreword

The 21st International Conference on Human-Computer Interaction, HCI International 2019, was held in Orlando, FL, USA, during July 26–31, 2019. The event incorporated the 18 thematic areas and affiliated conferences listed on the following page.

A total of 5,029 individuals from academia, research institutes, industry, and governmental agencies from 73 countries submitted contributions, and 1,274 papers and 209 posters were included in the pre-conference 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 pre-conference proceedings are listed in the following pages.

This year the HCI International (HCII) conference introduced the new option of "late-breaking work." This applies both for papers and posters and the corresponding volume(s) of the proceedings will be published just after the conference. Full papers will be included in the *HCII 2019 Late-Breaking Work Papers Proceedings* volume of the proceedings to be published in the Springer LNCS series, while poster extended abstracts will be included as short papers in the HCII 2019 *Late-Breaking Work Poster Extended Abstracts* volume to be published in the Springer CCIS series.

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 2019 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 2019 Constantine Stephanidis

HCI International 2019 Thematic Areas and Affiliated Conferences

Thematic areas:

- HCI 2019: Human-Computer Interaction
- HIMI 2019: Human Interface and the Management of Information

Affiliated conferences:

- EPCE 2019: 16th International Conference on Engineering Psychology and Cognitive Ergonomics
- UAHCI 2019: 13th International Conference on Universal Access in Human-Computer Interaction
- VAMR 2019: 11th International Conference on Virtual, Augmented and Mixed Reality
- CCD 2019: 11th International Conference on Cross-Cultural Design
- SCSM 2019: 11th International Conference on Social Computing and Social Media
- AC 2019: 13th International Conference on Augmented Cognition
- DHM 2019: 10th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management
- DUXU 2019: 8th International Conference on Design, User Experience, and Usability
- DAPI 2019: 7th International Conference on Distributed, Ambient and Pervasive Interactions
- HCIBGO 2019: 6th International Conference on HCI in Business, Government and Organizations
- LCT 2019: 6th International Conference on Learning and Collaboration Technologies
- ITAP 2019: 5th International Conference on Human Aspects of IT for the Aged Population
- HCI-CPT 2019: First International Conference on HCI for Cybersecurity, Privacy and Trust
- HCI-Games 2019: First International Conference on HCI in Games
- MobiTAS 2019: First International Conference on HCI in Mobility, Transport, and Automotive Systems
- AIS 2019: First International Conference on Adaptive Instructional Systems

Pre-conference Proceedings Volumes Full List

1. LNCS 11566, Human-Computer Interaction: Perspectives on Design (Part I), edited by Masaaki Kurosu
2. LNCS 11567, Human-Computer Interaction: Recognition and Interaction Technologies (Part II), edited by Masaaki Kurosu
3. LNCS 11568, Human-Computer Interaction: Design Practice in Contemporary Societies (Part III), edited by Masaaki Kurosu
4. LNCS 11569, Human Interface and the Management of Information: Visual Information and Knowledge Management (Part I), edited by Sakae Yamamoto and Hirohiko Mori
5. LNCS 11570, Human Interface and the Management of Information: Information in Intelligent Systems (Part II), edited by Sakae Yamamoto and Hirohiko Mori
6. LNAI 11571, Engineering Psychology and Cognitive Ergonomics, edited by Don Harris
7. LNCS 11572, Universal Access in Human-Computer Interaction: Theory, Methods and Tools (Part I), edited by Margherita Antona and Constantine Stephanidis
8. LNCS 11573, Universal Access in Human-Computer Interaction: Multimodality and Assistive Environments (Part II), edited by Margherita Antona and Constantine Stephanidis
9. LNCS 11574, Virtual, Augmented and Mixed Reality: Multimodal Interaction (Part I), edited by Jessie Y. C. Chen and Gino Fragomeni
10. LNCS 11575, Virtual, Augmented and Mixed Reality: Applications and Case Studies (Part II), edited by Jessie Y. C. Chen and Gino Fragomeni
11. LNCS 11576, Cross-Cultural Design: Methods, Tools and User Experience (Part I), edited by P. L. Patrick Rau
12. LNCS 11577, Cross-Cultural Design: Culture and Society (Part II), edited by P. L. Patrick Rau
13. LNCS 11578, Social Computing and Social Media: Design, Human Behavior and Analytics (Part I), edited by Gabriele Meiselwitz
14. LNCS 11579, Social Computing and Social Media: Communication and Social Communities (Part II), edited by Gabriele Meiselwitz
15. LNAI 11580, Augmented Cognition, edited by Dylan D. Schmorrow and Cali M. Fidopiastis
16. LNCS 11581, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Human Body and Motion (Part I), edited by Vincent G. Duffy

17. LNCS 11582, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Healthcare Applications (Part II), edited by Vincent G. Duffy
18. LNCS 11583, Design, User Experience, and Usability: Design Philosophy and Theory (Part I), edited by Aaron Marcus and Wentao Wang
19. LNCS 11584, Design, User Experience, and Usability: User Experience in Advanced Technological Environments (Part II), edited by Aaron Marcus and Wentao Wang
20. LNCS 11585, Design, User Experience, and Usability: Application Domains (Part III), edited by Aaron Marcus and Wentao Wang
21. LNCS 11586, Design, User Experience, and Usability: Practice and Case Studies (Part IV), edited by Aaron Marcus and Wentao Wang
22. LNCS 11587, Distributed, Ambient and Pervasive Interactions, edited by Norbert Streitz and Shin'ichi Konomi
23. LNCS 11588, HCI in Business, Government and Organizations: eCommerce and Consumer Behavior (Part I), edited by Fiona Fui-Hoon Nah and Keng Siau
24. LNCS 11589, HCI in Business, Government and Organizations: Information Systems and Analytics (Part II), edited by Fiona Fui-Hoon Nah and Keng Siau
25. LNCS 11590, Learning and Collaboration Technologies: Designing Learning Experiences (Part I), edited by Panayiotis Zaphiris and Andri Ioannou
26. LNCS 11591, Learning and Collaboration Technologies: Ubiquitous and Virtual Environments for Learning and Collaboration (Part II), edited by Panayiotis Zaphiris and Andri Ioannou
27. LNCS 11592, Human Aspects of IT for the Aged Population: Design for the Elderly and Technology Acceptance (Part I), edited by Jia Zhou and Gavriel Salvendy
28. LNCS 11593, Human Aspects of IT for the Aged Population: Social Media, Games and Assistive Environments (Part II), edited by Jia Zhou and Gavriel Salvendy
29. LNCS 11594, HCI for Cybersecurity, Privacy and Trust, edited by Abbas Moallem
30. LNCS 11595, HCI in Games, edited by Xiaowen Fang
31. LNCS 11596, HCI in Mobility, Transport, and Automotive Systems, edited by Heidi Krömker
32. LNCS 11597, Adaptive Instructional Systems, edited by Robert Sottilare and Jessica Schwarz
33. CCIS 1032, HCI International 2019 - Posters (Part I), edited by Constantine Stephanidis

34. CCIS 1033, HCI International 2019 - Posters (Part II), edited by Constantine Stephanidis
35. CCIS 1034, HCI International 2019 - Posters (Part III), edited by Constantine Stephanidis

http://2019.hci.international/proceedings

16th International Conference on Engineering Psychology and Cognitive Ergonomics (EPCE 2019)

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

- Shan Fu, P.R. China
- Qin Gao, P.R. China
- Wen-Chin Li, UK
- Peng Liu, P.R. China
- Heikki Mansikka, United Arab Emirates
- Ling Rothrock, USA
- Axel Schulte, Germany
- Alex Stedmon, UK

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-2019.php

HCI International 2020

The 22nd International Conference on Human-Computer Interaction, HCI International 2020, will be held jointly with the affiliated conferences in Copenhagen, Denmark, at the Bella Center Copenhagen, July 19–24, 2020. It will cover a broad spectrum of themes related to HCI, 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://2020.hci.international/.

General Chair
Prof. Constantine Stephanidis
University of Crete and ICS-FORTH
Heraklion, Crete, Greece
E-mail: general_chair@hcii2020.org

http://2020.hci.international/

Contents

Mental Workload and Performance

Goals – Assumption – Interaction Steps (GAIS): A Practical Method to Determine a Quantitative Efficiency Benchmark for UX Interaction Design Concepts 3
Helmut Degen

A Study on Visual Workload Components: Effects of Component Combination and Scenario Complexity on Mental Workload in Maritime Operation Tasks 20
Ye Deng, Yuexuan Wang, Manrong She, Yijing Zhang, and Zhizhong Li

Work Fragmentation, Task Management Practices and Productivity in Individual Knowledge Work 29
Heljä Franssila

Modeling of Operator Performance for Human-in-the-loop Power Systems . . . 39
Wan-Lin Hu, Claudio Rivetta, Erin MacDonald, and David P. Chassin

Comparing Eye-Gaze Metrics of Mental Workload in Monitoring Process Plants 55
Wenyan Huang, Yunshu Xu, Michael Hildebrandt, and Nathan Lau

Roster and Air Traffic Controller's Situation Awareness 66
Peter Kearney, Wen-Chin Li, and Graham Braithwaite

Socio-Technical Safety Investigations in Healthcare – Investigating Human Performance in Modern High Reliability Sector Organizations 76
Pete McCarthy and Andrew Blackie

Research on Workload-Based Prediction and Evaluation Model in Power System 92
Caifang Peng, Zhen Wang, Yanyu Lu, and Shan Fu

Development of a Quantitative Evaluation Tool of Cognitive Workload in Field Studies Through Eye Tracking 106
Giovanni Pignoni and Sashidharan Komandur

Towards a Mixed Reality Platform for Applied Cognitive Load Evaluation . . . 123
Maurice van Beurden and Linsey Roijendijk

Impacts of Automation Reliability and Failure Modes on Operators' Performance in Security Screening 137
Zijian Yin, Pei-Luen Patrick Rau, and Zhizhong Li

How Task Level Factors Influence Controllers' Backup Behavior: The Mediating Role of Perceived Legitimacy and Anticipated Workload 150
Saisai Yu, Jingyu Zhang, and Xiaotian E

Visual Cognition

A Visual Cognition Test-Based Study on the Choice Blindness Persistence: Impacts of Positive Emotion and Picture Similarity 167
Huayan Huangfu, Yi Lu, and Shan Fu

Color Ergonomics Research in Harsh Environment Under Three Task Modes 176
Zhiyang Jiang and Wenjun Hou

Semi-automatic Aggregation of Multiple Models of Visual Attention for Model-Based User Interface Evaluation. 187
Dennis Knoop, Bertram Wortelen, and Marcus Behrendt

Research on Aesthetics Degree Optimization Model of Product Form 200
Ming Li, Jie Zhang, and Yiping Hou

Research on Evaluation of Product Image Design Elements Based on Eye Movement Signal 214
Wenjin Yang, Jianning Su, Kai Qiu, Xinxin Zhang, and Shutao Zhang

Measurement of Human Sitting Posture Dimensions Using Human Pressure Distribution 227
Chen Yue, Linghua Ran, and Hua Qin

Effect of Mental Fatigue on Visual Selective Attention 236
Qianxiang Zhou, Jiaxuan Li, and Zhongqi Liu

Cognitive Psychology in Aviation and Space

Analysis of Key Cognitive Factors in Space Teleoperation Task 249
Junpeng Guo, Yuqing Liu, Xiangjie Kong, Shihua Zhou, and Jin Yang

Enhancing Aviation Simulator de-Briefs Through the Integration of Student Eye Tracking: The Instructor's Perspective. 259
Julius Jakubowski and Wen-Chin Li

Conceptual Modeling of Risk Interactions for Flight Crew Errors in Unmanned Aerial System Operation 271
Yi Lu, Huayan Huangfu, Shuguang Zhang, and Shan Fu

Checklist and Alert Language: Impact on ESL Pilot Performance in Airline Operations 284
Dujuan B. Sevillian

Fixation Adjustment During the Landing Process and Its Relationship with Pilot Expertise and Landing Performance 308
Yanjin Sun, Jingyu Zhang, Han Qiao, Xianghong Sun, Ping Qian, and Yang Song

From Paper-Based Operational Procedures to Onboard Context-Sensitive Information System (OCSIS) for Commercial Aircrafts 321
Wei Tan and Guy A. Boy

Airworthiness Compliance Criteria in Ergonomic Design of Cursor Control Device for Civil Aircraft. 332
Lei Wu and Jian Xu

Study on Evaluation of Airline Pilot's Flight Violation Behaviors and Psychological Risk 342
Jingyi Zhang and Lei Wang

Study of NASA-TLX and Eye Blink Rates Both in Flight Simulator and Flight Test 353
Yiyuan Zheng and Yuwen Jie

Group Collaboration and Decision Making

Spatial Ability in Military Human-Robot Interaction: A State-of-the-Art Assessment 363
Maartje Hidalgo, Lauren Reinerman-Jones, and Daniel Barber

Classification of Safety-Relevant Activities by Using Visual Scan Pattern in Airport Control Operations 381
Lothar Meyer, Åsa Svensson, Maximilian Peukert, Sven Malmberg Luengo, Jonas Lundberg, and Billy Josefsson

How Fire Risk Perception Impacts Evacuation Behavior: A Review of the Literature. 396
Hua Qin and Xiaotong Gao

Music Valence and Genre Influence Group Creativity 410
Hosseini Sarinasadat, Yuki Hattori, Yoshihiro Miyake, and Takayuki Nozawa

Human Operator Authentication Using Limited Voice Data: A Power Grid Dispatcher Instance. 423
Zheng Wang, Zhen Wang, Yanyu Lu, and Shan Fu

Correction to: Impacts of Automation Reliability and Failure Modes on Operators' Performance in Security Screening C1
Zijian Yin, Pei-Luen Patrick Rau, and Zhizhong Li

Author Index 439

Mental Workload and Performance

Goals – Assumption – Interaction Steps (GAIS): A Practical Method to Determine a Quantitative Efficiency Benchmark for UX Interaction Design Concepts

Helmut Degen(✉)

Siemens Corporation, Corporate Technology,
755 College Road E, Princeton, NJ 08540, USA
helmut.degen@siemens.com

Abstract. One reason for the digitalization megatrend is to further increase the efficiency of digital value chains. The human-machine interaction is typically one of the slowest elements in this chain, and therefore on the critical efficiency path. An efficiency increase of the human-machine interaction can lead to an efficiency increase of the entire digital value chain. To determine the efficiency of a UX design concept, an efficiency benchmark is needed before a UX design concept is created. The paper introduces GAIS (Goal – Assumption – Interaction Steps), a cognitive task analysis method. The author developed the GAIS method for use in industrial projects and validated it in industrial projects. GAIS allows to determine an interaction efficiency benchmark for a given use case. The benchmark is measured in number of interaction steps per use case. An important step in the GAIS method is a systematic reduction of interaction steps by considering manual, semi-automated and automated interaction. An interaction designer can use the determined efficiency benchmark to make an interaction design concept more efficient. The paper describes how GAIS is applied, with several examples (GUI and VUI), and how it guides the interaction design process.

Keywords: Cognitive task analysis · User experience efficiency benchmark · Multi-modality

1 Introduction

In today's world, the relevance and value of user experience (UX) design is more and more acknowledged. This is also true for industrial environments like Siemens. In industrial environments, one important UX quality is efficiency, the focus of this paper.

An interaction designer produces interaction design concepts, often in the form of wire frames and screen flows for Graphical User Interfaces (GUIs). Frequent questions are: How efficient is a proposed interaction design concept? How more efficient could the interaction design concept be? The questions are not only relevant for the interaction designer, but also for project stakeholders which like to know the quality of the presented UX work.

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 3–19, 2019.
https://doi.org/10.1007/978-3-030-22507-0_1

The question this paper attempts to answer is: How to determine a user experience efficiency benchmark? If such a user experience efficiency benchmark exists, it could be used to check the gap between the actual efficiency of an interaction design concept and such an efficiency benchmark.

This paper introduces GAIS (Goal - Assumption –Interaction Steps). The purpose of GAIS is to determine a user experience efficiency benchmark for a given use case which can be used to guide the design and the selection of interaction design concepts. GAIS has been successfully applied in several Siemens UX projects.

The paper is structured in the following way: In Sect. 2, GAIS is compared to GOMS. In Sect. 3, we discuss the efficiency benchmark dilemma. It motivates why an early efficiency benchmark is not a trivial endeavor. In Sect. 4, we explore the UX project context to explain when GAIS should be applied in a typical UX project. In Sect. 5, the GAIS approach will be presented, illustrated with several examples in Sect. 6 and lessons learned using GAIS in industrial projects. Section 7 concludes the use of GAIS and outlines future research and extension opportunities.

2 Related Work

GAIS is a distant relative of GOMS [1–3]. The purpose of GOMS (Goals, Operators, Methods, Selection rules) is to provide engineering models of human performance for human-computer interaction [1]. GOMS aims to optimize four criteria [4]:

C1 (a priori prediction) – a priori prediction of the human performance;
C2 (learnability and usability) – it can be used by computer system designers (not necessarily by psychologists or human factor experts).
C3 (coverage) – it covers a range of activities from perceptual-motor actions to complex activities to creative problem solving.
C4 (approximation) – it includes just the level of detail which is necessary for the design task. The purpose of using one of the GOMS derivates is to predict the human performance while interaction with a computer device.

The purpose of using one of the GOMS derivates is to predict the human performance while interaction with a computer device. GOMS breaks down the task hierarchy from goals to single interaction steps on several layers (goals and subgoals, operators, methods). GAIS has a different purpose, that of guiding rather than predicting efficiency. Instead of GOMS three aspect layers GAIS has two: Assumptions and Goals; and Interaction Steps. The reason for the simplicity is to consider only the necessary interaction steps in GAIS, and to ignore interaction steps which are the result of an interaction design process. GOMS considers such design decisions because it aims to make predictions about the outcome of the design process; GAIS does not considers such design decisions because it provides input for the design process and guides it. The design result may or may not follow identified GAIS options. GOMS is applied by computer system designers, GAIS is intended to be applied by human factor experts, psychologists, user experience designers, interaction designers etc., and not by technical experts.

3 Digitalization

3.1 UX System in a Digital Value Chain

A new megatrend Digitalization is “the use of digital technologies to change a business model and provide new revenue and value-producing opportunities” [6]. Two of the envisioned benefits of digitalization are speed and scale of processes [5, 7]. The focus of this papers is on speed. When we consider a human-in-the-loop, speed relates to how efficient a human can perform a task with the use of a machine.

The user is part of the digital value chain. A digital value chain is a process designed to create an outcome of value and which consists of digital tools, and users to create such an outcome. Typically, a user initiates the digital value chain. It is followed by automated steps which often require user involvement at some point. The result of involved users (user-machine interaction) and automated steps is an outcome of value (see Fig. 1).

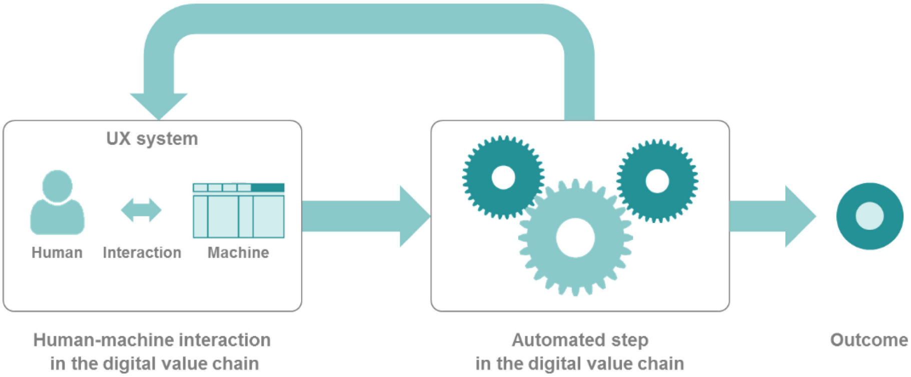

UX system
Human
Interaction
Machine
Human-machine interaction in the digital value chain
Automated step in the digital value chain
Outcome

Fig. 1. Digital value chain

Digital value chains in industrial environments are supported by digital tools, such as engineering tools, dashboards, configuration tools, HMIs of industrial devices, field service applications, IT applications etc.

Often, the least efficient part of such a digital value chain is the human-machine interaction (here also called the “UX system”). The human is a human actor, directly or indirectly interacting with a machine to produce the outcome of value. The machine is a human-made artifact designed to support producing the outcome.

If the performance of such an interaction can be increased, the performance of the entire digital value chain increases. That’s the reason why the UX system is on the critical efficiency path of a digital value chain.

Therefore, it is worthwhile to design the UX system with efficiency in mind. Efficiency, of course, is not the only relevant UX quality criterion, but an important one (see [8, 9] for other UX quality criteria such as effectiveness, satisfaction, accessibility, safety, well-being, and sustainability).

3.2 Interaction Types of UX Systems

Since the efficiency of the UX system is on the critical efficiency path, it is worthwhile to look further into interaction type which have a direct relevance for efficiency levels of the UX system. Without references to existing research results, the author introduces three interaction types.

Manual: The user interacts manually with the machine. There is no, or a very low degree of, system support. The efficiency level of the UX system is low.

Semi-automated: The machine takes over some of the manual interaction steps and or the machine provides recommendations to users, so the user saves interaction steps. In this case, the user is still in the loop and makes the final decision. The efficiency of the UX system is medium.

Automated: The user does not interact at all with the machine. All activities are performed by the machine. The level of automation is maximized. This is the highest level of efficiency. The efficiency of the UX system is high.

The three interaction types and their differences are illustrated with an example for entering data into a form and using the data:

- Manual: The user enters data manually into a form (Interaction step 1); the user reviews and submits the data (interaction step 2).
- Semi-automated: The machine populated the data automatically into the form, based on past data entries (no user interaction). The user reviews the data and submits the data (interaction step 1).
- Automated: The machine fetches automatically the data from past data usage (no interaction step) and submits the data automatically to where they are needed (no interaction step). The user is not involved at all.

All three interaction types are relative to the state-of-the-art of interaction technologies. What we consider “manual” interaction today might be “semi-automated” yesterday. Here is an example how to create a shape in a vector graphics tool: Today, we draw a shape with a mouse or pen. We would categorize this user task it as a “manual” interaction type. Before the first GUI-based system was invented, around 40 years ago, the user needed to define a shape by specifying its coordinates. At that time, the definition of a shape with coordinates would be considered “manual” interaction type and using a mouse (or a pen) and a GUI would be considered “semi-automated”.

When applying GAIS for a given use case, the question will arise whether a manual interaction type can be made more efficient by introducing a semi-automated or even an automated interaction type. A technical solution exists in many, but not in all cases for introducing a semi-automated or automated interaction types.

It is known that semi-automated and automated interaction types have their own challenges. Some of them are described in [10, 11]. GAIS does not consider those automation related challenges. They need to be considered and evaluated separately.

3.3 Research Question

The motivating question for this research effort was: How to determine a user experience efficiency benchmark? The efficiency benchmark is intended to be used to determine the actual efficiency of a drafted interaction design concept, and to determine the efficiency gap between the interaction design concept and the efficiency benchmark.

If such a method is expected to be accepted in an industrial environment, additional requirements apply:

- R1 (Core): For a given use case, the method shall be used to determine a quantitative UX efficiency benchmark. Note: The requirement articulates the core outcome of the method which is the determination of a UX efficiency benchmark.
- R2 (Independency): The method shall create an efficiency benchmark independent from the creation of an interaction design concept. Note: The method needs to be allocated to a step in the UX process prior to creation of interaction design concepts.
- R3 (Scalability): The method shall be applied to larger UX elements (e.g. an entire UX framework) as well as to single UX elements (e.g. a single UI widget). Note: The method should not be constraint to a smaller (widget level) or larger (framework level) UX scope.
- R4 (Modality): The method shall be applicable to different interaction modalities, or combinations of them, such as Direct Manipulation, Voice, Gesture. Note: This requirement addresses the trend that more and more UX solutions include more than one interaction modalities. Examples of interaction modalities are: Direct Manipulation (with GUIs), Voice User Interfaces, Gesture etc.
- R5 (Effort): For a given use case, the method shall allow an interaction designer to create an efficiency benchmark in equal to or less than one working day. Note: The cost/benefit ratio needs to be reasonable to be applicable in industrial projects. “One working day” is an arbitrary, but reasonable upper limit for use in industrial projects.
- R6 (Explainability): The method shall allow UX stakeholders to understand the efficiency benchmark and how to apply it. Note: The underlying assumption for this requirement is that a UX professional should be able to explain and defend a proposed UX solution. The method to determine a UX benchmark should therefore be explainable to UX stakeholders. (Subjective) feedback from UX stakeholders help to check whether this requirement is met.

This paper proposes GAIS as such a method. Before GAIS is introduced, additional information about the UX process performed at Siemens Corporate Technology is presented.

4 User Experience Background

4.1 User Experience Process

At Siemens Corporation, Corporate Technology (CT), we apply the user experience process, depicted in Fig. 2. The process consists of four phases.

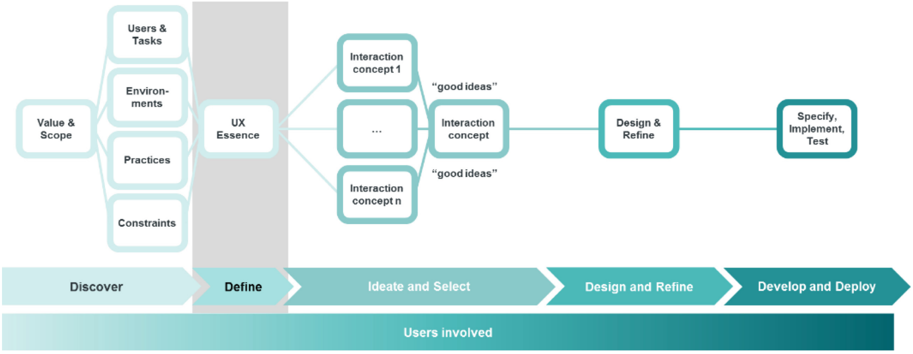


Fig. 2. User experience process; the efficiency benchmark should be identified during the "Define" step and document as part of "UX Essence"

Discover: In an initial "Value and Scope" phase, the UX team understands the business case of the product under design and how UX can support the business case. It also outlines the initial UX scope. Afterwards, the UX team gains an understanding about the involved user roles and their user profiles (e.g. use cases, user needs, user characteristics), the use environments (e.g. spatial, work flow, social), relevant good and bad practices and known constraints (i.e. business, technology, design, and regulatory).

Define: The insights are consolidated in a step called "UX Essence" which includes UX goals, optimization use cases, and UX quality and quantity criteria which are used to access explored interaction design concepts.

Ideate and Select: The UX teams create several interaction design concepts and assess them against the established UX quality and quantity criteria. If the UX criteria are not met, further interaction design concept options will be created. The UX team finally settles on an interaction design concept which ideally meets all the defined UX criteria.

Design and Refine: The UX teams refine the selected interaction design concept, adds missing details and creates the visual design. Users are involved to evaluate the designs. UX designers refine the design according user's feedback.

Develop and Deploy: The UX teams creates optionally a specification or another kind of document as input for the front-end development and implements the front-end. The implementation can be a prototype or a product-quality front-end. The UX teams may evaluate the implemented prototype/front-end (e.g. with usability tests) and refines the design afterwards to address the findings.

Some additional explanations about the outlined UX process:

1. Representatives from identified user groups and project stakeholders (e.g. product manager which determines the business case and selects product features; project manager which keeps the project on track in terms of quality, time, and cost; front- and back-end developers which implement the UX and technical design) are involved in all phases, so close feedback loops are happening along the way. For instance, we prefer involvement of user reps on a weekly basis. If users have

concerns, the process may go a step back, e.g. from “Ideate and Select” to “Learn and Define”, or from “Design and Refine” to “Ideate and Select”. These loops are not displayed in Fig. 2.

2. This UX process can be applied to an entire UX framework (e.g. an Engineering Tool) as well as to a single UX element (e.g. Find and Replace widget). For an entire UX framework, more time is necessary for each phase than when the process is applied to a single UX element.
3. This UX process can be applied to agile, waterfall, or hybrid (called “wagile”) product development approaches. It is critical that the first phases (“Discover”, “Define”, “Ideate and Select” and “Design and Refine”) are performed before the UX results are implemented.

Back to the efficiency topic: When should the UX efficiency benchmark be determined? Since the efficiency is determined for use cases, it can be applied only after the use cases are identified. This means it can only be applied after the “Discover” phase is completed. The efficiency benchmark should be used as input into the creation of interaction design concepts. Therefore, the benchmark should be determined before the “Ideate & Select” phase. Therefore, the appropriate phase is the “Define” phase (see highlight in the Fig. 2).

5 GAIS Method

5.1 Elements

The GAIS method uses the following structural elements: Goal, Assumptions, and Interaction Steps. The way to visualize the result of a GAIS analysis can vary. One way to visualize the structural elements of GAI it is shown in Fig. 3.

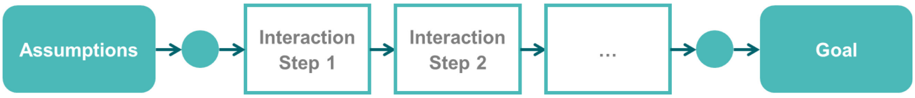


Fig. 3. GAIS elements

Here a few explanations:

- The GAIS method is applied to a single use case.
- A goal is an intended state for a given use case.
- Assumptions are an initial state before the identified interaction steps are performed to achieve the goal.
- The identified interaction steps are the result of a cognitive task analysis for the given use case.

- The identified interaction steps should consider necessary steps for the given use case, and they should not consider design decisions or design elements (e.g. select an item from a list). Otherwise, unnecessary design decision may over constraint the design space.
- The identified interaction steps should consider design constraints. These are steps which must be executed, due to business or regulatory constraints.
- The interaction steps should consider the interaction nature of the selected modality or modalities.
- The granularity of interaction steps per GAIS analysis can vary. For instance, for a low-level use case (e.g. identify payment information), the GAIS interaction steps can relate to single mouse clicks (for a direct manipulation interaction modality); for a high-level use case (e.g. order a product), the interaction steps may relate to a sequence of mouse clicks. It is assumed that a single GAIS analysis identifies a similar granularity level for the identified interaction steps for a given use case.

5.2 Measurement Units

It is useful to introduce some language which explains how efficient an interaction design concept is, compared to the determined efficiency benchmark. The envisioned number of interaction steps for a use case is called the "Benchmark Interaction Steps" (abbreviated "BIS"). The "Actual Interaction Steps" (abbreviated "AIS") is determined by counting the actual number of interaction steps of an interaction design concept to perform a given use case.

To determine the efficiency category of an interaction design concept, we use the following terminology:

- If AIS > BIS: the interaction design concept is called "inefficient"
- If AIS = BIS: the interaction design concept is called "efficient"
- If AIS < BIS: the interaction design concept is called "super-efficient"

The BIS is determined during the "Define" phase, and the AIS is determined during the "Explore and Select" phase (see Fig. 2).

5.3 GAIS Process

The GAIS method is applied to a single use case. The GAIS process consists of four steps (see Fig. 4). For a given use case, the interaction designer identifies the goal (intended state) (step 1). In addition, the designer defines the assumptions, which are the pre-conditions or initial state before the user performs any interaction (step 2). Afterwards, the interaction designer outlines the interaction steps for a first GAIS option, typically the "manual" efficiency type (step 3). Afterwards, the interaction designers may identify a semi-automated interaction option and fully automated interaction option (step 4), where applicable. The interaction designer may iterate step 3 and 4 (Fig. 4).

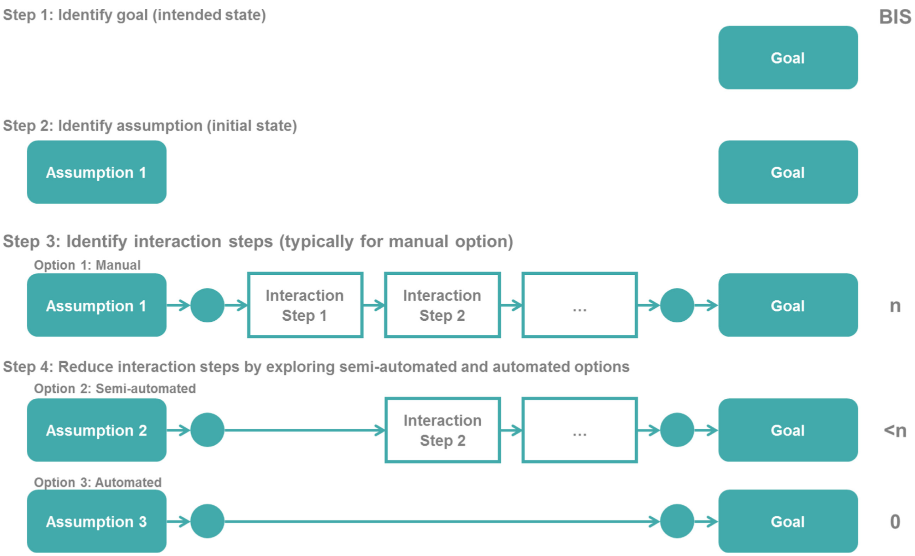

Step 1: Identify goal (intended state)
BIS
Goal
Step 2: Identify assumption (initial state)
Assumption 1
Goal
Step 3: Identify interaction steps (typically for manual option)
Option 1: Manual
Assumption 1
Interaction Step 1
Interaction Step 2
...
Goal
n
Step 4: Reduce interaction steps by exploring semi-automated and automated options
Option 2: Semi-automated
Assumption 2
Interaction Step 2
...
Goal
<n
Option 3: Automated
Assumption 3
Goal
0

Fig. 4. GAIS process

To achieve efficiency improvements with a semi-automated or automated GAIS option, single interaction steps are assumed to be performed automatically by the machine. The result of that performance is added to the assumption box (see Fig. 5).

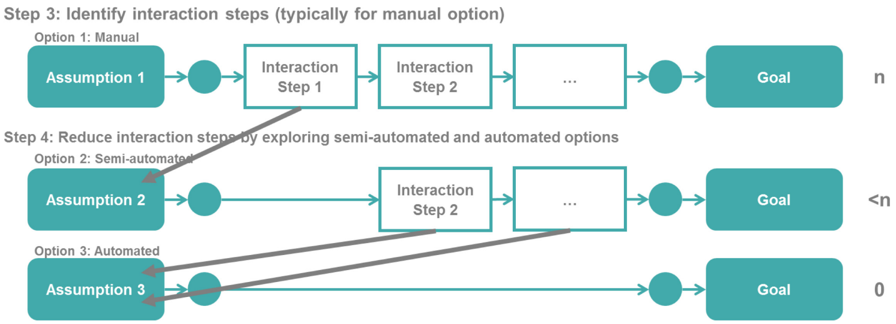


Fig. 5. GAIS process; results of interaction steps are added to assumptions.

Afterwards, the interaction designer can make notes of the BIS for each identified GAIS option by counting the number of interaction steps. When creating an interaction design concept, the designer can then compare the different BIS with the AIS of the actual interaction design concept and can iterate the interaction design concept until the AIS is equal to or even better than the BIS. Let's look at some examples.

6 Examples

6.1 Example 1: Order a Product

The first example illustrates how GAIS can be used to check whether an existing design is inefficient, efficient or super-efficient. We picked an everyday use case “Order a product”. Three different BISs are depicted in Fig. 6.

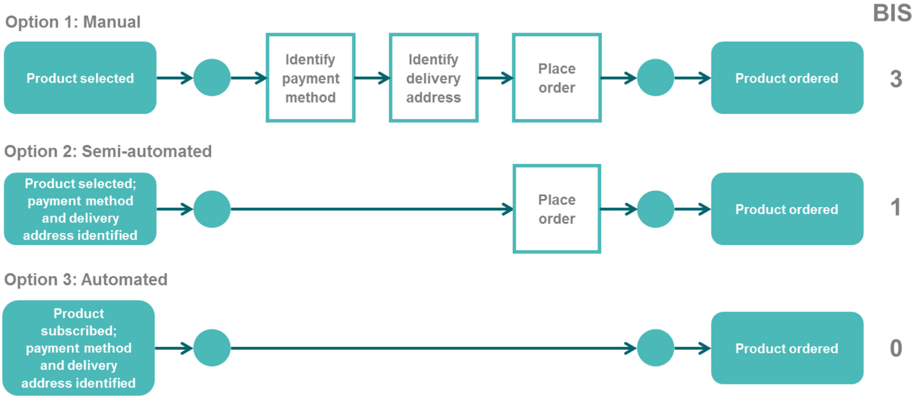


Fig. 6. GAIS, applied to use case “Order a product”

For all three GAIS options, it is assumed that the product is already selected (as stated in the Assumption box). Option 1 shows the manual option. It reflects that a user identifies the payment method (interaction step 1), the shipping address (interaction step 2) and then places the order (interaction step 3). The BIS is therefore 3.

For the use case “Order a product”, the manual interaction can be optimized by considering a semi-automated approach which is shown in option 2. Since the order process is often applied many times, the optimization applies to the identification of the payment and shipping information. The efficiency optimization semi-automates the use of the payment und shipping information. Every time the user wants to order a product, the order and shipping information is automatically fetched, so the user does not have to enter it manually. This leads to a changed assumption: product selected, payment information identified, shipping information identified. The only step the user must perform is to order the product. The BIS went down from 3 steps (option 1) to 1 step (option 2).

Option 3 is further optimized the use case by identifying a fully automated option. In this case, even placing the order is automated. This might be useful if a user knows that s/he needs to order a certain product frequently. In such a case, automating the process might be more convenient than a manual order. This is reflected in option 3. The payment method and the delivery address are identified, including an order schedule. The product will automatically be ordered and shipped following the defined order schedule. The BIS went down from 1 (option 2) to 0 (option 3).

This example reflects what Amazon™ has implemented. Option 1 reflects ordering a product manually, by manually entering the payment method and the shipping information the first time (AIS: 3). Option 2 reflects the 1-Click™ order which Amazon™ has implemented and patented [12] (AIS: 1). Option 3 reflects Amazon's™ subscription feature (AIS: 0). If we compare the three BISs with the AIS of Amazon's™ implementation, then the Amazon™ implementation is "GAIS efficient" in all three cases.

Amazon's™ Dash Button™ is very interesting. It selects and orders a product with one interaction steps (AIS: 1), literally with one click. Option 2 assumes that a product is already selected. The selection of a product requires at least one interaction step. The Dash Button™ does both with one interaction step. Therefore, the Dash Button™ is "super-efficient".

6.2 Example 2: Respond to an Alarm

The second example illustrates how GAIS can guide the design of an interaction design concept for an industrial example. The modality is GUI. The use case is that an operator responds to a received alert (Fig. 7).

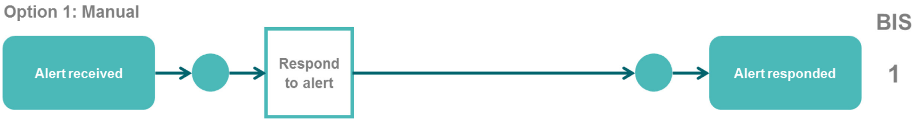


Fig. 7. GAIS, applied to use case "Respond to alert"

In this case is only one GAIS option with a BIS of 1 (see Fig. 7). The reason for the one step is a business constraint which does not allow to make it more efficient. In the UX project, different wire frame options with different Actual Interaction Steps (AIS) were created (see Fig. 8).

In the first GAIS option, the user receives a notification (interaction step 1). After the user selected the notification, the user is taken to a list of alerts. The user selects the alert on the top of the list (interaction step 2). The user is taken to a detailed view. The user reads the description of the alert and responds with populated options (interaction step 3). The AIS of option 1 is 3 ("Inefficient" because AIS > BIS).

In the second GAIS option, the information of the alert summary (in the alert list) is moved over to the notification. After the user has received the notification, the user selects the notification is taken directly to the alert description (interaction step 1). After the user has read the alert description, the user responds with populated options (interaction step 2). The AIS is option 2 is 2 ("inefficient" because AIS > BIS).

In the third GAIS option, the alert description and the populated options are displayed in the notification. The user reads the alert and response to it with populated options. The AIS of option 3 is 1 ("efficient" because AIS = BIS) (Fig. 8).

The example illustrates that the BIS, determined with the GAIS method, can guide the refinement of the interaction design concept to achieve systematically a higher efficiency. It also shows that the interaction designer when to stop making the

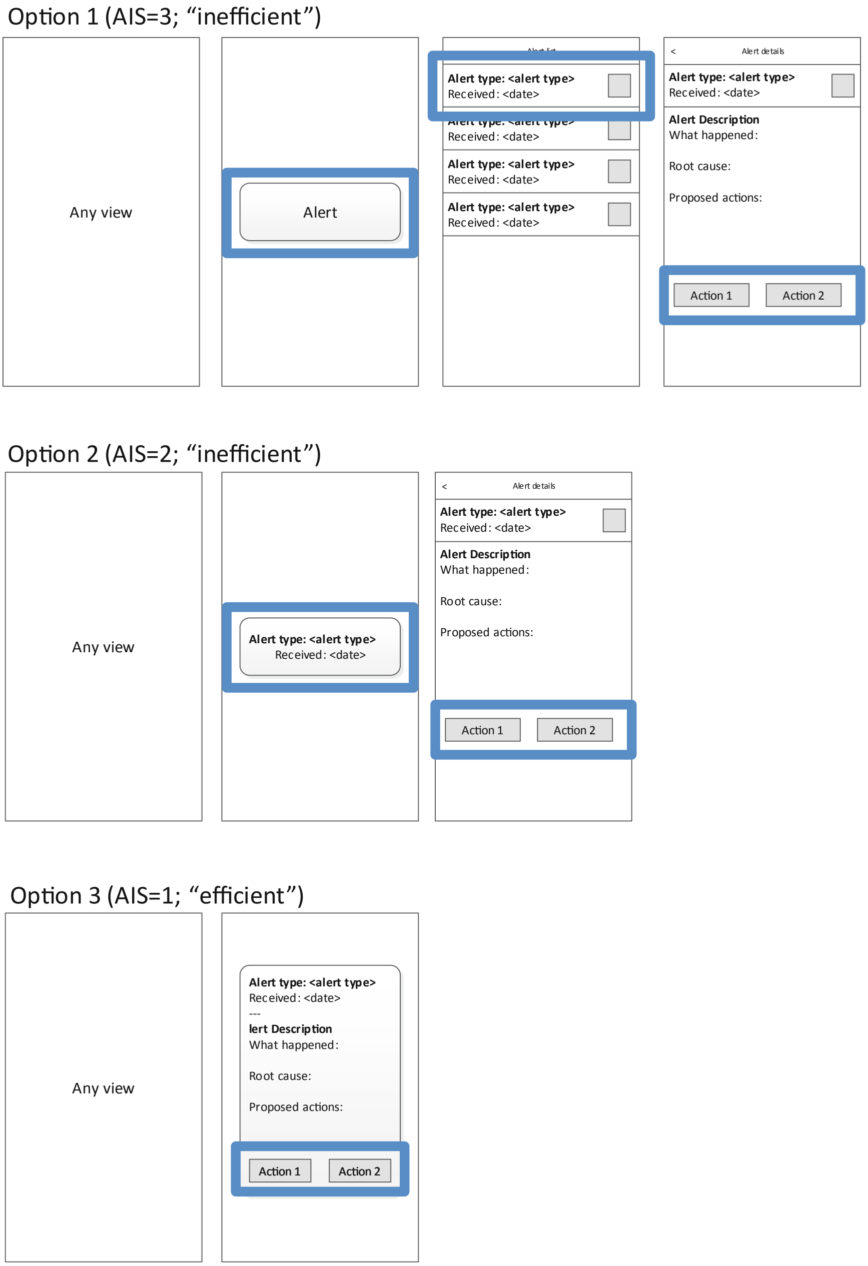


Fig. 8. Wire frame options for use case "Respond to alert"

interaction design concept more efficient. As this example, and the other examples illustrate, the BIS provides an objective measure for efficiency, however it is determined by the interaction designer and therefore subjective.

6.3 Example 3: Create an Engineering Artifact

A third example illustrates how GAIS can be applied to an engineering tool. The chosen use case is “Create engineering artifact”. The modality is GUI. The first GAIS option shows the creation of an engineering artifact, step by step. The BIS is n which equals the number of engineering elements added. The semi-automated option reduces it to one interaction step, under the assumption that reusable templates for such an engineering artifact are available (see Fig. 9).

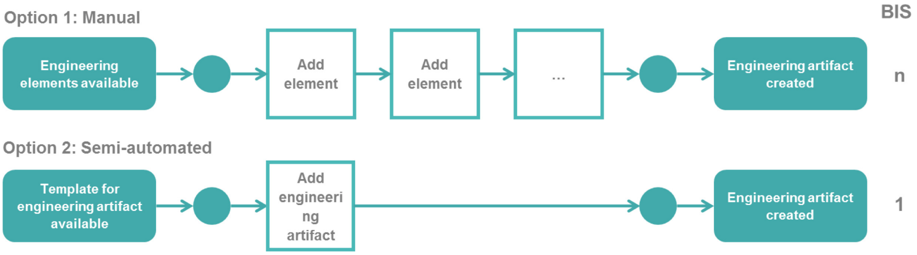

Option 1: Manual
BIS
Engineering elements available
Add element
Add element
...
Engineering artifact created
n
Option 2: Semi-automated
Template for engineering artifact available
Add engineeri ng artifact
Engineering artifact created
1

Fig. 9. GAIS applied to use case “Create engineering artifact”

The GAIS method can also be applied to the use case “Create engineering artifact template” with two options (see Fig. 10).

In option 1, the user creates the templates manually, and in option 2, the engineering tool automatically creates templates for established engineering artifacts for later reuse.

As shown in this example, the usefulness of GAIS lies in the systematic consideration of efficiency improvements, considering semi-automated and automated interaction types (Fig. 10).

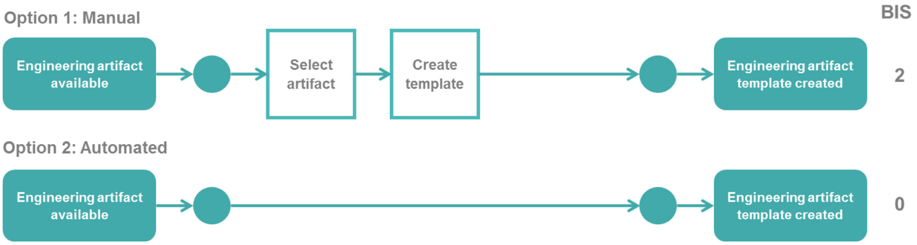

Option 1: Manual
BIS
Engineering artifact available
Select artifact
Create template
Engineering artifact template created
2
Option 2: Automated
Engineering artifact available
Engineering artifact template created
0

Fig. 10. GAIS applied to use case “Create engineering artifact template”

6.4 Example 4: Change Motor Speed

This example illustrates how different interaction modalities influence the BIS. In the third example, we compare how the speed of a motor can be changed with a Graphical User Interface (GUI) and a Voice User Interface (VUI).

The first GAIS option (see Fig. 11) shows two interaction steps for a GUI: Select a motor and set the speed of the motor (BIS: 2). The second GAIS option shows the same use case for a VUI. It has only one step because the user can utter in one sentence "Change the speed of motor 1 to 250 rpm", and the VUI is assumed to understand both parameters [13]. This is often not possible with a GUI where interaction is typically serialized. Beside hands-free usage, the efficiency increase, compared to GUIs, is another advantage of VUIs. This increase is reflected in the BIS (=1) for option 2, compared to a BIS (=2) for option 1 (Fig. 11).

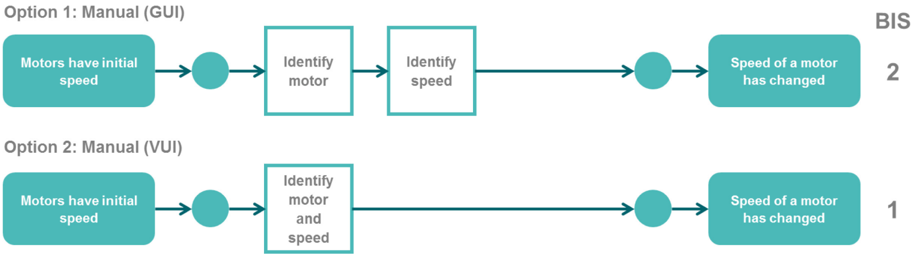

Option 1: Manual (GUI)
BIS
Motors have initial speed
Identify motor
Identify speed
Speed of a motor has changed
2
Option 2: Manual (VUI)
Motors have initial speed
Identify motor and speed
Speed of a motor has changed
1

Fig. 11. GAIS applied to use case "Change motor speed" (GUI and VUI)

6.5 Lessons Learned from Using GAIS

One question is to which use cases GAIS should be applied. GAIS can be applied to all use cases (the author does it) of an UX project, however if the reader wants to try it out, it is recommended to consider the following types of use cases:

- High frequency use cases (the efficiency improvements have the biggest impact for high frequency use cases)
- Urgency (these use cases do often not occur frequently, but when they occur, they are optimized for efficiency)

The highest efficiency improvements lie in finding semi-automated and automated interaction options. This means that additional functionality needs to be developed. If the semi-automated or automated options are preferred by users, the author encourages UX people to request the needed additional functionality. UX does not stop at the visualization layer.

The author also experienced that users sometimes do not want semi-automated or automated support because it takes away control from users. Loss of control needs to be taken seriously. Therefore, it is important to involve users when envisioning semi-automated or automated interaction options. Users should provide feedback whether they accept a proposed semi-automated or automated option, or not. If users are concerned about an automated option, try to understand the underlying concern. Sometimes, a settlement on a semi-automated option is agreeable, keeping users in the loop and let the user make the final decision. In general, user acceptance is more important than efficiency.

A designer applying GAIS may face the situation that the user preferences are split. If the development budget allows the implementation of several options, the designer may want to go for it. If development budget is only available for one option, you may pick the one which satisfy most of the critical users.

GAIS identifies an efficiency benchmark which is used as an objective measure. However, the determination of a GAIS option is the result of a cognitive task analysis, performed by a user experience designer and still subjective.

In this paper, the GAIS options are visualized with diagrams. Other ways to visualize the GAIS options are possible, e.g. as a bulleted list. It could look like this:

- Assumption: Motors has initial speed
  - Interaction step 1: Identify motor
  - Interaction step 2: Identify speed
- Goal: Motor speed has changed

A bulleted list is easier to create and maintain than a diagram. The diagram has advantages when it comes to comparing different GAIS options and recognizing their differences. The preferred way to visualize depends on the audience. The author used the diagram technique with a positive feedback from project stakeholders. They understood the intent of the method and the resulting BIS per GAIS option.

Another visualization option is to make explicit which interaction steps were added to assumptions. This can be done as illustrated in Fig. 5.

7 Conclusion

GAIS is a cognitive task analysis method which allows the UX designer to identify several interaction steps, the so-called benchmark interaction steps (BIS). The BIS is used to inform and guide the subsequent design of interaction design concepts. When creating an interaction design concept, the actual number of interaction steps (AIS) can be determined and compared to the BIS. If the number of actual interaction steps (AIS) for a given use case is larger than BIS, then the interaction designer has a hint to refine the interaction design concept. With GAIS, an interaction designer can make an interaction design concept systematically more efficient by comparing the AIS with the BIS and exploring semi-automated and automated interactions. By comparing the AIS with the BIS, the interaction designer knows how efficient an interaction design concept is.

The intent of GAIS is not to predict the efficiency or effort, like GOMS does, but to identify an efficiency benchmark. As shown in this paper, GAIS can be applied to different modalities, e.g. direct manipulation, voice or multimodal interfaces.

Let's check whether GAIS complies with the six requirements:

- R1 (Core): GAIS supports the determination of the number of interaction steps for a given use case, including different options. The number of interaction steps becomes the efficiency benchmark.

- R2 (Independency): GAIS is applied during the “Define” phase which takes place before the interaction design concepts are created during the “Explore and Select” phase.
- R3 (Scalability): This requirement refers to the scope of the use case. The use case can be very broad, so an entire UX framework is needed for the implementation and the determination of the AIS. Example 1 is such broad use case. The use case can also have a smaller scope and will be implemented by a single UX widget. Example 4 is such a use case. GAIS can be used for both cases.
- R4 (Modality): GAIS is agnostic of interaction modalities. However, when applying GAIS, it should be made explicit which interaction modalities is used because the GAIS option can lead to a different BIS, depending on the selected modality. This was demonstrated with example 4 (VUI vs. GUI).
- R5 (Effort): In our experience, when applying GAIS, it usually only takes a few minutes to find different options and determine the BISs.
- R6 (Explainability): GAIS options, their BISs and the corresponding interaction design concepts with their AISs were presented in several projects to project stakeholders. They could easily follow the approach and the thought process. GAIS helps to make design decision explainable and defendable.

Our conclusion is that GAIS complies with all six requirements.

GAIS has limitations, though. The GAIS method does not consider other user experience qualities, like learnability, safety, accessibility etc. In other words: the application of GAIS does not guarantee a usable design. That means, the interaction design, informed by a BIS, need to go through the normal, overall UX process, including a usability evaluation. This is particularly important when semi-automated or automated options are proposed. User acceptance is still more important than efficiency. In addition, GAIS does not consider effort for performing single interaction steps. This is kept out deliberately as a simplification. However, it is a potential area for extension. Finally, research is needed to extend the GAIS method to address other user experience qualities.

Acknowledgement. I thank Christof Budnik and Michael Little for constructive feedback.

References

1. Card, S.K., Moran, T.P., Newell, A.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, London (1983)
2. John, B.E., Kieras, D.E.: The GOMS family of analysis techniques: tool for design and evaluation. CMU-CS-94-181, 24 August 1994
3. Kieras, D.E.: GOMS Models for task analysis. In: Diaper, D., Stanton, N. (eds.) The Handbook of Task Analysis for Human-Computer Interaction. Lawrence Erlbaum, London (2004)

4. 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)
5. Siemens: The race to a digital future. Assessing digital intensity in US manufacturing. https://www.siemens.com/content/dam/internet/siemens-com/us/home/company/topic-areas/digitalization/documents/cg-mc-digital-future-of-us-manufacturing-en.pdf. Accessed 16 Oct 2018
6. Gartner IT Glossary: Digitalization. https://www.gartner.com/it-glossary/digitalization/. Accessed 16 Oct 2018
7. Roland Busch: Innovation at speed and scale. https://www.siemens.com/press/pool/de/events/2017/corporate/2017-12-innovation/presentation-roland-busch-innovation.pdf. Accessed 16 Oct 2018
8. ISO 9241-110:2006: Ergonomics of human-system interaction – Part 110: Dialog principles
9. ISO 9241-210:2010: Ergonomics of human–system interaction — Part 210: Human-centred design for interactive systems
10. Onnasch, L., Wickens, C.D., Li, H., Manzey, D.: Human performance consequences of stages and levels of automation: an integrated meta-analysis. Hum. Factors: J. Hum. Factors Ergonomics Society **56**(3), 476–488 (2013)
11. Kaber, D.B., Endsley, M.R.: The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. Theor. Issues Ergonomics Sci. **5**(2), 113–153 (2004)
12. Hartman, P., Bezos, J.P., Kaphan, S., Spiegel, J.: Method and system for placing a purchase order via a communications network. United States Patent and Trademark Office US 5,960,411, 28 Sept 1999
13. Pearl, C.: Designing Voice User Interfaces. O’Reilly Media, Sebastopol (2017)

A Study on Visual Workload Components: Effects of Component Combination and Scenario Complexity on Mental Workload in Maritime Operation Tasks

Ye Deng[1], Yuexuan Wang[2], Manrong She[2], Yijing Zhang[2], and Zhizhong Li[2](✉)

[1] China Institute of Marine Technology & Economy, Beijing, People's Republic of China
dengye@cimtec.net.cn

[2] Department of Industrial Engineering, Tsinghua University, Beijing 100084, People's Republic of China
wyx_thu14@qq.com, {shemanrong, ldlzyj}@163.com, zzli@tsinghua.edu.cn

Abstract. Mental workload is a key factor in influencing human performance of maritime tasks among which visual tasks dominate. Through a lab experiment, this study examined the effects of visual workload components, their combinations and scenario complexity on mental workload. 20 participants were recruited to perform visual tasks on a simulated platform. Results showed that some combinations of visual components significantly increased the participants' mental workload. Scenario complexity was found to have significant interaction effects with workload component combinations. The results indicate that because of their negative effects on task performance, some combinations of workload components should be specially avoided in task design.

Keywords: Mental workload · Visual task component · Scenario complexity · Maritime operation

1 Introduction

In safety-critical industries, mental workload is considered as a key factor for operator performance [1]. Based on the theory of resource demand and supply, mental workload is defined as the relationship between the mental resources demanded by a task and those resources to be supplied by an operator [2]. In a review of maritime accident reports, researchers illustrated how the 13 out of 31 accidents was caused by either too high or too low level of mental workload [3]. During the system design, usability testing and operator training, it is important not only to measure the level of mental workload, but also to predict how much mental workload an operator would perceive while performing the tasks [4].

The Visual/Auditory/Cognitive/Psychomotor (VACP) model is a popular method to predict mental workload. VACP was first developed for a military lightweight

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 20–28, 2019.
https://doi.org/10.1007/978-3-030-22507-0_2

helicopter system [5]. To adopt the VACP method, researchers first break the mission into phases, segments, functions, tasks, and the performance elements for each task. A performance element is usually composed of a verb and an object. On the basis of multiple resource theory, workload components are categorized into four channels (i.e. visual, auditory, cognitive and psychomotor channels). Through expert judgment, a standard rating scale for workload components was proposed. For example, the component "detect" in the visual channel is rated 1.0 (for the detailed rating scale, please refer to [4, 5]. For each performance element, researchers identify its relevant workload components and sum up the scores by each channel. In this way, the workload for each task is quantified in terms of the four channels. The score is determined by the nature of the task, and is independent of individual factors.

The VACP method is not without potential flaws. As we know, mental workload is influenced not only by the demand/resource balance, but also by other factors such as time pressure, scenario complexity, individual experience and ability [6, 7]. Under a high complexity scenario (e.g. more targets on the display) or under high time pressure, an operator may perceive higher mental workload even if they are executing the same task as before [8]. In addition, when using the scale, the scores are sum up for a specified time period, without considering whether the workload components in this period are presented in series or parallel, meaning that combination of workload components (multi-task) is not differentiated from individual workload components. In the current study, we provided a list of visual workload components for maritime operation tasks, and investigated whether different components and their combinations would influence operator's mental workload, and whether the influence would change under different scenario complexity levels.

2 Visual Workload Components for Maritime Operations

Through mission segmentation and expert judgment, we revised McCracken and Aldrich's workload components [5] to make them applicable for maritime operation tasks. The process of developing maritime workload components would not be detailed in this paper. We focused on the workload components from the visual channel. The visual workload components are shown in Table 1. Two components in Mccracken and Adrich's model, "scan/search/monitor" and "read", were replaced by "retrieve" and "compare" as described in Table 1.

3 Experiment

The experiment was conducted to investigate whether different workload components and combinations of them would influence operator's mental workload, and whether the influence would be different under different scenario complexities. The maritime operation tasks were designed to meet the study requirements of different workload components. Participants executed the tasks using a simulated maritime platform. Their performance data were recorded for later analysis.

Table 1. Description of visual workload components.

Workload component	Description
Retrieve	Look through the information, retrieve the needed information, obtain the parameter values
Compare	Compare different pieces of information visually to select the proper one(s)
Detect	Detect the appearance and disappearance of an object in the field of vision
Discriminate	Discriminate the change (e.g. color, shape, value, location) of an object
Search	Search a certain object in the field of vision
Check	Check the state of equipment/parameter, check the action/command to be performed
Track	Track/follow an object in order to find out any abnormalities

3.1 Participants

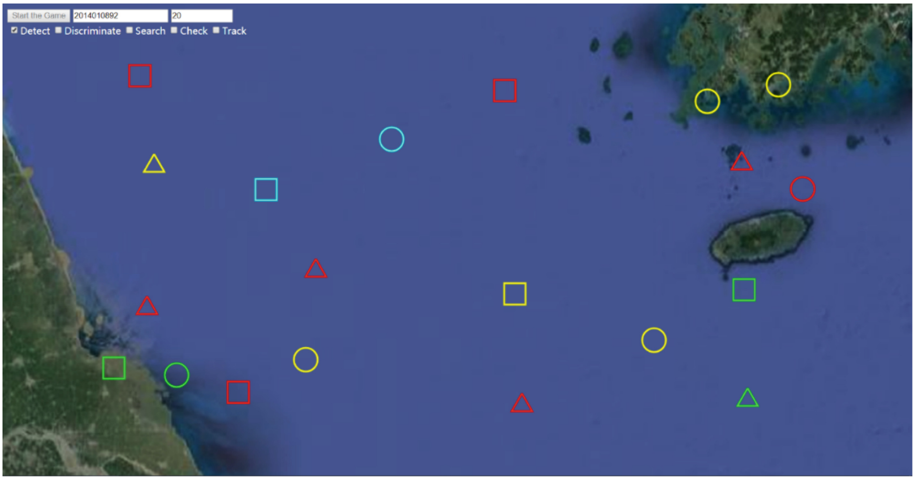
Twenty undergraduates from Tsinghua University (10 males and 10 females) were recruited as participants. Their average age was 20 with the standard deviation of 2.8. All the participants had normal or correct-to-normal vision. The participants were informed of the experimental details and voluntarily signed the consent form.

3.2 Experimental Platform

A simulated maritime operation system was developed. The interface is shown in Fig. 1. The targets were represented with different colors (i.e. red, blue, yellow and green) and different shapes (i.e. square, circle and triangle). The system log data and the participant's performance data were recorded and exported to text files.

3.3 Independent Variables

There were two independent variables in the experiment: scenario complexity and combination of workload components. Both of them were within-subject variables. Scenario complexity was defined as the number of targets on the maritime display. The three levels of scenario complexity were low (10 targets on the display), medium (20 targets on the display), and high (30 targets on the display). The combination of workload components included detect, discriminate, search, check, track, and the combination of any two of above. The combination of search and check were excluded because the "search" and the "check" tasks could not be executed at the same time. Therefore, there were 14 combinations of workload components. Each participant was required to perform tasks in all cases. A case meant a single workload component or the combination of two workload components combined with a level of scenario complexity. The participant was supposed to repeat the task for 10 trials in a case.



2014010892 20

☑ Detect □ Discriminate □ Search □ Check □ Track

Fig. 1. The maritime interface (Color figure online)

3.4 Tasks and Procedure

At the beginning of the experiment, the participants filled out the demographic information. The experimenter introduced the purpose of the experiment, the tasks to be performed, and the usage of the simulated platform. The participants then practiced with the platform for about 10 min to get familiar with the whole experiment. The experimenter would answer any questions during the practice.

Before the experiment, the participant adjusted the seat height and the distance to the computer screen. Every time the participant completed the task in a case, he/she rested for 15 s.

During the formal experiment, the experimenter first entered a scenario complexity level in the system, and the display would present the corresponding number of targets to the participant. The tasks to be performed were as follows.

- Detect. Within the time limit, new targets randomly appeared on the display. The participant was expected to detect the newly appeared target and click on it as quickly as possible. If the participant did not respond to the new target within 5 s, the system would record it as time-out.
- Discriminate. Within the time limit, an existing target on the display changed its color. The participant was asked to click on the target as soon as he/she noticed the color change. If the participant did not respond to the target within 5 s, the system would record it as time-out.
- Search. The system randomly presented searching-relevant questions for the participant to answer.
- Check. The interface was divided into six areas. The participant was asked to decide whether the total number of targets within an area exceeded a certain value.

- Track. A new target appeared and randomly moved on the display. If the new target ran into an exiting one, the participant was asked to click on the existing target that was ran into. If the participant did not respond within 5 s, the system would record it as time-out.

The error rate and the response time were recorded for all the above tasks.

3.5 Dependent Variables

Performance data could reflect the level of mental workload. Although workload is not the only factor that influences operator performance [9], higher workload is believed to contribute to worse performance [6]. In the experiment, the error rate and the response time were recorded as two dependent variables [10].

4 Results

4.1 Effects of Single Workload Component and Scenario Complexity

Table 2 summarizes the descriptive statistics of single workload component and scenario complexity on error rate and response time.

Table 2. Descriptive statistics of single workload component and scenario complexity.

		Error rate			Response time		
Workload component	Scenario complexity	*N*	Mean	SD	*N*	Mean (ms)	SD (ms)
Detect	Low	20	0.065	0.0813	20	1288	147.8
	Medium	20	0.120	0.1281	20	1327	221.9
	High	20	0.185	0.1461	20	1270	246.8
Discriminate	Low	20	0.300	0.1654	20	1450	307.0
	Medium	20	0.450	0.2013	20	1436	245.9
	High	20	0.520	0.1361	20	1487	260.6
Search	Low	20	0.040	0.0821	20	4365	1183
	Medium	20	0.050	0.1100	20	6573	1521
	High	20	0.110	0.1373	20	8994	2923
Check	Low	20	0.017	0.0745	20	1574	492
	Medium	20	0.025	0.0816	20	1667	698
	High	20	0.033	0.1026	20	1924	774
Track	Low	20	0.015	0.0489	20	808	254.4
	Medium	20	0.040	0.1146	20	773	183.0
	High	20	0.035	0.0587	20	838	247.2

The error rate data violated the assumption of normality. The nonparametric Kruskal-Wallis test was used. The effect of scenario complexity on error rate was significant with the “detect” ($p = 0.020$) and “discriminate” ($p < 0.001$) tasks, but was not significant with the “search” (p = 0.267), “check” (p = 0.954), and “track” (p = 0.559) tasks. For the “detect” and “discriminate” tasks, the higher scenario complexity resulted in higher error rate. Under the same scenario complexity, the error rates among different workload components were significantly different (all $p < 0.001$). Under the medium scenario complexity level (i.e. 20 targets on the display), Mann-Whitney’s U test was used for post hoc analysis. The post hoc results are shown in Table 3. The error rate exhibited a significant difference between the “detect” task and any of the other four tasks. The same could be said between the “discriminate” task and any of the other four tasks. The error rates between two of the “search”, “check”, and “track” tasks were not significantly different.

The natural logarithm of the response time data satisfied the normality and homogeneity assumptions, and thus ANOVA was used. The main effects of workload component (F = 64.70, $p < 0.001$) and scenario complexity (F = 3.73, p = 0.025) were significant. The interaction effect between workload component and scenario complexity was also significant (F = 3.73, p = 0.025). Tukey’s method was used for multiple comparisons. Results showed that the differences between low and high scenario complexity levels and between medium and high scenario complexity levels were significant. Except for between “detect” and “discriminate” and between “discriminate” and “check”, the differences between any other two of the workload components were significant.

Table 3. Post Hoc analysis (Mann-Whitney’s U Test) for error rate

Workload component 1	Workload component 2	W	p
Detect	Discriminate	248.50	<0.001
Detect	Search	483.00	0.0499
Detect	Check	506.00	0.010
Detect	Track	488.00	0.036
Discriminate	Search	587.50	<0.001
Discriminate	Check	593.00	<0.001
Discriminate	Track	587.50	<0.001
Search	Check	431.00	0.579
Search	Track	406.00	0.925
Check	Track	383.00	0.473

4.2 Effects of Workload Component Combinations and Scenario Complexity

In this section, we compared the differences between one workload component and its combinations with the other components. For example, the difference between the “detect” task and the combination of “detect” and “discriminate” tasks were examined in terms of error rate and response time. The detailed results are illustrated as follows.

The "Detect" Task and its Combinations

The results of Kruskal-Wallis test showed that under the medium (p = 0.006) and high (p = 0.002) scene complexities, the error rates between the "detect" task and its combinations with other tasks were significantly different. Under the high scenario complexity, the results of Mann-Whitney test showed that compared with the single "detect" task, adding the "search" task significantly increased the participant's error rate (p = 0.019). Adding the other tasks did not increase the error rate significantly.

In terms of response time, the results of ANOVA showed that the combinations of workload components had a significant effect ($p < 0.001$). The effects of scenario complexity (p = 0.982) and their interaction (p = 0.792) were not significant. Post hoc analysis revealed that compared with the "detect" task, adding the "search" ($p < 0.001$) or "check" ($p < 0.001$) tasks significantly increased the participant's response time.

The "Discriminate" Task and its Combinations

In terms of error rate, the results of Kruska-Wallis test did not show any difference between the "discriminate" task and its combinations with other tasks.

Compared with the single "discriminate" task, adding other tasks did not increase or decrease the participant's response time obviously.

The "Search" Task and its Combinations

The results of Kruskal-Wallis test revealed that the error rate between the "search" task and its combinations with other tasks were different under the medium (p = 0.005) and the high (p = 0.017) scenario complexity levels. Further analysis (Manny-Whitney test) showed adding the task of "detect" or "discriminate" significantly increased the participant's error rate.

As for response time, the effects of workload component combination and scenario complexity, and their interaction effect were significant. The response time was significantly increased when the "detect" ($p < 0.001$), "discriminate" ($p < 0.001$), or "track" ($p < 0.001$) task was added to the "search" task.

The "Check" Task and its Combinations

No matter under which level of scenario complexity, error rate did not exhibit difference between the "check" task and its combinations with other tasks, according to the results of Kruskal-Wallis test.

We only analyzed 14 participants' response data due to data missing. The ANOVA results showed that different combinations of workload components had a significant effect on response time ($p < 0.001$). The effect of scenario complexity (p = 0.904) and the interaction effect (p = 0.206) were not significant. By adding the "detect" ($p < 0.001$), "discriminate" (p = 0.024), or "track" ($p < 0.001$) task, the response time was significantly increased.

The "Track" Task and its Combinations

The results of Kruskal-Wallis test showed that no matter under which scenario complexity level, the error rates were not different between the "track" task and its combinations with other tasks.

In terms of response time, there were no significant effects of workload component combinations (p = 0.347), scenario complexity (p = 0.446), or their interaction (p = 0.706).

5 Discussion

According to the statistics of "mental workload" in the publication title from Ergo-Abs database, the mental workload research in maritime engineering seems not as active as those in driving and air-traffic control [11]. In this study, an experiment was conducted to investigate the effects of visual workload components on operator's mental workload inferred from task performance, and whether the effects would be different if the scenario complexity varied. Operator's mental workload was evaluated through two performance measures: error rate and response time.

Results showed that some combinations of workload components significantly increased the participant's mental workload (e.g. higher error rate, more response time). Compared with the single "detect" or "search" task, the combination of the two tasks led to higher workload. In another case, adding the "track" task on the basis of the "search" or "check" task obviously increased the workload. However, adding the "search" or "check" to the "track" task was not found to make a difference in error rate or response time. The possible reason is that the "track" task takes up a larger proportion of the visual resources than the "search" or "check" task and requires the participant's continuous attention. If operator performance is essential for the system effectiveness and safety, the workload component combinations that would degrade performance should be avoided during task design. If the combination cannot be avoided in real working environment, other channels (e.g. the auditory channel) should be used to release the workload.

The VACP rating scale provides standard values for workload components. However, scenario complexity was found to have interaction effects with workload component combinations. How much workload a workload component brings might be affected by the scenario complexity. In this case, the standard value given by the scale may not reflect the accurate workload that an operator would perceive. Further investigation is needed to make clear whether and how other factors interact with workload components to affect mental workload.

The study has several limitations. This study provided evidence that scenario complexity should be considered in the VACP component rating, but did not indicate how the scenario complexity should be quantified in the scale. Due to the limitation of the designed simulated system, the "comparison" task and the combination of "search" and "check" tasks were not included in the experiment. Moreover, the study focused on the single workload component and the combination of any two. It needs further investigation on whether and how the combinations of more than two components would influence task performance and mental workload.

6 Conclusion

This study reveals that workload components would lead to different workloads when they are combined, and they would interact with scenario complexity to influence operator's mental workload. Based on the experimental results, we provide some suggestions on task design and workload component scaling.

Acknowledgement. This study was supported by the projects with the Grant No. JCKY2016 206A001 and No. 6141B03020604.

References

1. Wickens, C.D.: Multiple resources an: mental workload. Hum. Factors **50**, 449–455 (2008)
2. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cognitive Eng. Decis. Making **2**, 140–160 (2008)
3. Embrey, D., Blackett, C., Marsden, P., Peachey, J.: Development of a Human Cognitive Workload Assessment Tool. Human Reliability Associates, Dalton (2006)
4. Aldrich, T.B., Szabo, S.M., Bierbaum, C.R.: The development and application of models to predict operator workload during system design. In: McMillan, G.R., Beevis, D., Salas, E., Strub, M.H., Sutton, R., Van Breda, L. (eds.) Applications of Human Performance Models to System Design, pp. 65–80. Springer, Boston (1989). https://doi.org/10.1007/978-1-4757-9244-7_5
5. McCracken, J.H., Aldrich, T.B.: Analyses of selected LHX mission functions: implications for operator workload and system automation goals (No. ASI479-024-84). Anacapa Sciences Inc., Fort Rucker AL (1984)
6. Vidulich, M.A., Tsang, P.S.: Mental workload and situation awareness. In: Salvendy, G. (ed.) Handbook of Human Factors and Ergonomics, vol. 4, pp. 243–273. (2012)
7. Wu, Y., Miwa, T., Uchida, M.: Using physiological signals to measure operator's mental workload in shipping – an engine room simulator study. J. Marine Eng. Technol. **16**, 61–69 (2017)
8. 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)
9. Wickens, C.D., Hollands, J.G., Banbury, S., Parasuraman, R.: Engineering Psychology & Human Performance. Psychology Press, London (2015)
10. Mazur, L.M., Mosaly, P.R., Hoyle, L.M., Jones, E.L., Marks, L.B.: Subjective and objective quantification of physician's workload and performance during radiation therapy planning tasks. Pract. Radiat. Oncol. **3**, 171–177 (2013)
11. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics **58**, 1–17 (2015)

Work Fragmentation, Task Management Practices and Productivity in Individual Knowledge Work

Heljä Franssila(✉)

Faculty of Information Technology and Communication Sciences (ITC),
Tampere University, 33014 Tampere, Finland
helja.franssila@gmail.com

Abstract. The study examined the nature of task management practices, their prevalence and relations to experiences of work fragmentation and productivity among Finnish state governmental organization employees, all of them knowledge workers. In the descriptive analysis it was found out, that knowledge workers experience most often work fragmentation as experiences of extreme hurry and forgetfulness. When considering productivity, respondents were more often satisfied with the quality of work they were able to fulfil, but less often to the amount of work they were able to finish. Less than half of the respondents collect and list all of their tasks into one place regularly. Every fifth of respondents newer plan or write down work duties and goals for the beginning work week, and every fourth of respondents never decide the start and due date for their single work tasks. Nearly every fifth never utilised any digital tool to support any personal task management activity. The correlation analysis revealed that negative correlation between the experiences of work fragmentation and productivity was statistically significant. Experiences of effectiveness of task management was negatively correlated with work fragmentation. Finally, maturity of applied task management practices was positively correlated with effectiveness of task management.

Keywords: Work fragmentation · Task management · Productivity · Knowledge work

1 Introduction

Maintaining sense of coherence and satisfying levels of personal productivity in a daily basis is a challenge in hectic contemporary knowledge work life. Maintenance of smooth work flow and ability to concentrate fully in the work duties is hard. Experiences of performance losses caused by the fragmentation of work are common in contemporary knowledge work settings. Task execution and getting tasks fulfilled is hampered by both external and internal interruptions, creating a workflow containing constant and rapid task switching (Czerwinski et al. 2004; Iqbal and Horvitz 2007). Considerable share of external distractions originate from digital work environment (Franssila et al. 2014). Conventional approach to support knowledge workers to maintain more productive workflow and avoid work fragmentation has concentrated on

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 29–38, 2019.
https://doi.org/10.1007/978-3-030-22507-0_3

different ways interface design and application settings can protect task execution from external distractions (e.g. Iqbal and Horvitz 2007; Lindlbauer et al. 2016). However, a considerable share of interruptions are self-generated (Dabbish et al. 2011; Adler and Benbunan-Fich 2013). Instead, very little in known about daily task management and task execution planning practices of knowledge workers.

In order to better understand conditions and consequences of task management and task planning in individual knowledge work, this study examined the nature of task management practices, their prevalence and relations to experiences of work fragmentation and productivity among Finnish state governmental organization knowledge workers. The research questions of this study were:

(1) What is the nature of work fragmentation experiences in knowledge work settings?
(2) How work fragmentation is related to the experiences of task management effectiveness and personal productivity?
(3) What is the nature and prevalence of task management practices applied by knowledge workers?
(4) How the maturity of task management practices and experiences of task management effectiveness are related?
(5) How the maturity of task management practices and experiences of work fragmentation and personal productivity are related?

2 Background

Despite widespread experiences of hurry, work fragmentation and interruptions in knowledge work settings, surprisingly small amount of academic research has empirically observed knowledge worker task management and task execution planning practices (Haraty et al. 2016). In the academic literature, concepts of task management, task planning, activity planning and time management have been applied somewhat interchangeably, referring to broad categories of activities related to task planning and task scheduling (see Claessens et al. 2009a). Several studies examine the complicated and overloaded role email often plays in practical task management of knowledge workers (Bellotti et al. 2005; Whittaker et al. 2007). In an interview and video-diary study observing academic professionals it was found out, that the main task management activities applied were planning, prioritization and list-making. Task management and task execution contained various challenges. In particular in task management, several difficulties were experienced: integrating different media, rearranging tasks, determining appropriate tasks, identifying reasonable timeframes, generating flexible schedules, managing long term goals, estimating task duration and differentiating the nature of different tasks. Considering actual task performance, another set of difficulties were identified: accomplishing competing tasks, undertaking planned tasks, undertaking long term goals, undertaking tasks that do not involve other people, remembering small tasks, retaining self-motivation and assessing previous achievements (Kamsin 2014). However, it was left unclear, how actively and regularly academics executed different task management activities. In another diary and survey study actual task completion of R&D engineers was monitored. In the multi level analysis it was found out, that the tasks with higher priority, urgency and lower importance were more likely to be completed. The time management training reveiced recently was one of the individual level explainers of the task completion rate, alongside with the personality trait of conscientiousness and emotional stability. However, self-reported inclination to planning was not related to higher rate of actual task completion (Claessens et al. 2009b). On the other hand, time management training does not always quarantee actual application of learned strategies into daily work. In a study observing impact of time management training into perceived stress, perceived control of time and performance at work, large differences in the actual application of learned strategies were found (Häfner and Stock 2010).

According the theory of implementation intention and goal attainment (Gollwitzer and Oettingen 2016), when one has explicitly specified when, where and with which sub-steps one is going to fulfil a task goal, it has tendency to become fulfilled. The basic script of specifying implementation intentions resembles the main elements of task management and task planning – deriving sub-goals and concrete task from main goals, estimating time required to complete different tasks, understanding temporal interdependencies between tasks, putting various task into order of execution, scheduling both long and short term task execution and monitoring the rate of task accomplishment. It can be hypothesized, that in knowledge work settings which are filled with variety of goals and variety of possibilities to organize one's duties, specifying implementation intentions and "scripting" one's goal attainment may enhance sense of coherence and personal productivity, and finally even eliminate the stressful experiences of work fragmentation. On the other hand, can the nature and maturity level (or lack) of task management methods and practices applied explain at least a share of concurrent experiences of work fragmentation? While several studies have evaluated and given design recommendation for specific digital tool functionalities to support task management, the overall understanding of core processes and applied practices of task management in the real world among knowledge workers has remained vague. This study provides description of the task management practices applied in knowledge work settings and provides preliminary evidence that development and deployment of task management skills can enhance knowledge worker productivity.

3 Methodology

The data of the study was collected with an online survey distributed to knowledge workers (n = 59) employed in a governmental organization in autumn 2018. The survey was delivered to a group of volunteer participants in the organization. Variable amount of expert, management and support duties were included into the work roles of the survey participants. Most of the participants (95%) had expert duties, 24% of participants had managerial duties and 22% had support duties in their task profile.

The survey contained measures to assess as dependent variables daily experiences of work fragmentation and personal productivity developed in Franssila et al. (2016). Respondents were asked to indicate their experiences in overall during the last five working days. As independent variables were measures of qualitative nature and effectiveness of applied task management practices. Measures for task management practices were designed for the purpose of this study, operationalizing the theories of implementation intention and goal striving (Gollwitzer and Oettingen 2016).

Measures of work fragmentation, productivity and task management effectiveness were composed of item statements, and respondents were asked to assess the item statements on 5-point rating scale, from "I strongly disagree" (=1) to "I strongly agree" (=5) (see Table 1). The final measures of work fragmentation, task management effectiveness and productivity were calculated by summing the scores of item variables according to Table 1. The reliabilities of measures were evaluated with Cronbach's alpha, and they were the following: work fragmentation (α = 0,75), task management effectiveness (α = 0,88) and productivity (α = 0,49).

Table 1. Description of the survey measures – work fragmentation, task management effectiveness and productivity.

Measure	Items in the measure (rating scale: 1 = I strongly disagree–5 = I strongly agree.)
	Considering my personal work performance during the last five work days, I was experiencing
Work fragmentation	Intensive hurry Forgetfulness Too frequent disruptive interruptions Difficulties to concentrate to tasks at hand Difficulties to complete tasks which I believed I could complete today
Task management effectiveness	Ease of prioritization of my daily work Ease of deciding the task execution order in my daily work
Productivity	Satisfaction with the quality of completed work Satisfaction with the amount of completed work

Measure of task management practices was composed of item statements considering application of various task management practices, and respondents were asked to assess the item statements on 3-point rating scale, with scores "I apply this practice regularly" (=1), "I apply this practice time to time" (=2), and "I never apply this practice" (=3) (see Table 2). The final measure of task management practices was calculated by summing the scores of item variables according Table 2. The reliability of the measure of task management practices was evaluated with Cronbach's alpha, and the score was α = 0,75.

Table 2. Description of the survey measures – task management practices.

Measure	Items in the measure (rating scale: "I apply this practice regularly" (=1), "I apply this practice time to time" (=2), and "I never apply this practice" (=3)
Task management practices	I collect and list all my tasks into one place I "take an inventory" of my tasks regularly by checking with tasks are completed and which are uncompleted I organize and split goals of my work role and tasks into concrete subtasks I classify my tasks according to which goals and responsibility areas of my work role they serve I examine my work task load as a whole in order to see how different goal areas of my work role are represented there I classify my tasks according importance I classify my tasks according attractiveness I classify my tasks according urgency I budget (= estimate and book) time for different tasks and for different responsibility areas of my work role I plan and record tasks and goals for the beginning work week I plan and record tasks and goals for the beginning work day I decide when I start and complete certain work task I schedule all my tasks (not only meetings and appointments) to be completed in certain time

In addition, open ended questions were included into the survey. In open ended questions, respondents were able to describe in their own words, what kind of digital tools and practices they applied in their daily task management, if any. In particular, the tools and practices applied in listing, evaluating, screening and organizing their tasks and managing time were asked to be described in the responses. From qualitative, written responses the amount of mentions of different applications and tools were recorded.

In the analysis of survey data, first, descriptive statistics of measures were calculated. Next, correlation analysis to study relations between dependent and independent variables was executed. Correlations between work fragmentation, productivity experiences and the maturity level of task management practices were statistically tested.

4 Results

In the descriptive analysis it was found out, that knowledge workers experience most often work fragmentation as experiences of extreme hurry and forgetfulness, but a bit less often difficulties to complete the duties they has planned for the day (Table 3). When considering productivity, respondents were more often satisfied with the quality of work they were able to fulfil, but less often to the amount of work they were able to finish (Table 4). Prioritization and planning the execution order of the tasks were equally challenging task management activities (Table 5).

Table 3. Descriptive statistics of items of work fragmentation measure (n = 59).

Items (scale: 1 = I strongly disagree–5 = I strongly agree.)	Mean	SD
Intensive hurry	3,66	0,93
Forgetfulness	3,59	1,00
Too frequent disruptive interruptions	3,37	1,05
Difficulties to concentrate to tasks at hand	3,36	0,99
Difficulties to complete tasks which I believed I could complete today	3,11	1,19

Table 4. Descriptive statistics of items of productivity measure (n = 59).

Items (scale: 1 = I strongly disagree–5 = I strongly agree.)	Mean	SD
Satisfaction with the quality of completed work	3,83	0,56
Satisfaction with the amount of completed work	3,07	0,96

Table 5. Descriptive statistics of items of task management effectiveness measure (n = 59).

Items (scale: 1 = I strongly disagree–5 = I strongly agree.)	Mean	SD
Ease of prioritization of my daily work	3,19	0,86
Ease of deciding the task execution order in my daily work	3,14	0,86

Analysis of responses to open ended questions revealed, that variety of distinct digital applications and functionalities were utilized in one or several distinct task management activities. Wide variety of applications were utilized in note-taking related to the task management. Tools supporting paper-mimicking note-taking and reminders (MS Onenote and Sticky Notes) were rather actively applied. Also a share (15%) of respondents utilized the specific, integrated task management tool Outlook Tasks (Table 6).

Table 6. Variety of digital applications utilized in task management based on open responses.

Tool/application	% of users
MS Outlook (in general)	66
MS Outlook Calendar	53
MS Onenote (in personal use)	34
MS Sticky Notes	24
MS Outlook Tasks	15
Project management application	12
MS Excel	9
Categories in MS Outlook	7
Trello	5
MS Outlook Email	5
MS Word	5
Kanbanflow	5

(*continued*)

Table 6. (*continued*)

Tool/application	% of users
MS Planner	3
MS Onenote (in collaborative use)	2
MS Sharepoint collaboration site	2
Reminders in MS Outlook	2
Notepad	2
Windows Resource Manager	2

Nearly every fifth never utilised any digital tool to support any personal task management activity. Some of the participants utilized both paper-based and digital means to manage their tasks. In overall, it was not possible to determine, what other than digital tools were actually applied in recording and scheduling tasks. In addition, when respondent mentioned "Outlook" as an application they utilized, it is impossible to determine, which tool/tools of Outlook they were actually using.

Less than half of the respondents collect and list all of their tasks into one place regularly. Nearly every fifth of respondent newer plan or write down work duties and goals for the beginning workweek, and every fourth of respondents never decide the start and due date for their single work tasks (Table 7).

Table 7. Descriptive statistics of items of task management practices measure (n = 59).

Item	% respondents who *never* apply the practice	% respondents who *from time to time* apply the practice	% respondents who *regularly* apply the practice
I collect and list all my tasks into one place	7	47	46
I "take an inventory" of my tasks regularly by checking with tasks are completed and which are uncompleted	7	61	32
I organize and split goals of my work role and tasks into concrete subtasks	20	58	22
I classify my tasks according to which goals and responsibility areas of my work role they serve	64	28	9
I examine my work task load as a whole in order to see how different goal areas of my work role are represented there	53	44	3
I classify my tasks according importance	5	39	3

(*continued*)

Table 7. (*continued*)

I classify my tasks according attractiveness	55	41	56
I classify my tasks according urgency	0	29	3
I budget (=estimate and book) time for different tasks and for different responsibility areas of my work role	12	61	71
I plan and record tasks and goals for the beginning work week	21	53	26
I plan and record tasks and goals for the beginning work day	15	58	27
I decide when I start and complete certain work task	27	56	17
I schedule all my tasks (not only meetings and appointments) to be completed in certain time	14	66	20

In the correlation analysis of measures of work fragmentation, productivity, task management effectiveness and task management practices applied it was found out, that negative correlation between the experiences of work fragmentation and productivity was statistically significant (r = −0,430, p = 0,001). Experiences of effectiveness of task management was negatively correlated with work fragmentation (r = −0,451, p = 0,000). Finally, maturity of applied task management practices was positively correlated with effectiveness of task management (r = 0,299, p = 0,022).

5 Discussion

This study was one of the first academic examinations of prevalence of task management and task planning activities in real life knowledge work context. The results of the study show, that negative experiences of work fragmentation, loss of control over task execution and lost productivity are less common among knowledge workers who proactively manage their task load, and organize and plan their task execution. When considering countermeasures to hinder productivity losses related to interruption-prone contemporary work environments, more emphasis should be put into the development of task management practices and their training and implementation among knowledge workers. When most of the knowledge work assignments are both delivered and executed in digital work environment, the skills and practices of efficient digital task management and task execution planning are critical to enhance productivity and to mitigate stress and negative mental workload created by work fragmentation.

6 Limitations

Because the empirical data collected in this study was based on subjective assessments on frequency of task management activities, certain bias compared to the actual practices applied may exist. Another limitation of the study is the small size of the survey data and the inclusion of only one organization into analysis.

7 Conclusions

Despite widespread experiences of hurry, work fragmentation and interruptions in knowledge work settings, surprisingly small amount of academic research has empirically observed knowledge worker task management and task execution planning practices (Haraty et al. 2016). The results of this study show, that the more comprehensive the repertoire of task management practices actually applied in everyday work, the higher the experience of task management effectiveness. Further, the experience of effectiveness of task management was related to experiences of lower work fragmentation. Experiences of work fragmentation and personal productivity were related – the higher the experiences of fragmentation, the lower the experiences of personal productivity. While several studies have evaluated and given design recommendation for specific digital tool functionalities to support task management, the overall understanding of core processes and applied practices of task management in the real world among knowledge workers has remained vague. This study provided description of the task management practices applied in knowledge work settings and provides preliminary evidence that development and deployment of task management skills can enhance knowledge worker productivity.

Acknowledgements. This study would not have been possible without the helpful Human Resources staff of the participating governmental organization and the efforts of all the survey participants. The conference presentation of this paper in HCI International 2019 –conference was supported by Finnish Work Environment Fund.

References

Adler, R., Benbunan-Fich, R.: Self-interruptions in discretionary multitasking. Comput. Hum. Behav. **29**(4), 1441–1449 (2013)

Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., Grinter, R.E.: Quality versus quantity: E-mail-centric task management and its relation with overload. Hum.-Comput. Interact. **20**(1–2), 89–138 (2005)

Claessens, B.J., Roe, R.A., Rutte, C.G.: Time management: logic, effectiveness and challenges. In: Roe, R.A., Waller, M.J., Clegg, S.R. (eds.) Time in Organizational Research, pp. 23–41. Routledge, Abingdon (2009a)

Claessens, B.J., Van Eerde, W., Rutte, C.G., Roe, R.A.: Things to do today: a daily diary study on task completion at work. Appl. Psychol. **59**(2), 273–295 (2009b)

Czerwinski, M., Horvitz, E., Wilhite, S.: A diary study of task switching and interruptions. In: Proceedings of the CHI 2004, pp. 175–182. ACM (2004)

Dabbish, L., Mark, G., González, V.M.: Why do i keep interrupting myself? Environment, habit and self-interruption. In: Proceedings of the CHI 2011, pp. 3127–3130. ACM (2011)

Franssila, H., Okkonen, J., Savolainen, R.: Email intensity, productivity and control in the knowledge worker's performance on the desktop. In: Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services, pp. 19–22. ACM (2014)

Franssila, H., Okkonen, J., Savolainen, R.: Developing measures for information ergonomics in knowledge work. Ergonomics **59**(3), 435–448 (2016)

Gollwitzer, P.M., Oettingen, G.: Planning promotes goal striving. In: Vohs, K., Baumeister, R. (eds.) Handbook of Self-Regulation, pp. 223–244 (2016)

Haraty, M., McGrenere, J., Tang, C.: How personal task management differs across individuals. Int. J. Hum. Comput. Stud. **88**, 13–37 (2016)

Häfner, A., Stock, A.: Time management training and perceived control of time at work. J. Psychol. **144**(5), 429–447 (2010)

Iqbal, S.T., Horvitz, E.: Disruption and recovery of computing tasks: field study, analysis, and directions. In: Proceedings of the CHI 2007, pp. 677–686. ACM (2007)

Kamsin, A.: Improving tool support for personal task management (PTM). Doctoral dissertation, UCL, University College London (2014)

Lindlbauer, D., Klemen, L., Walter, R., Müller, J.: Influence of display transparency on background awareness and task performance. In: Proceedings of the CHI 2016, pp. 1705–1716 (2016)

Whittaker, S., Bellotti, V., Gwizdka, J.: Everything through Email. In: Personal Information Management, pp. 167–189 (2007). University of Washington Press

Modeling of Operator Performance for Human-in-the-loop Power Systems

Wan-Lin Hu[1,2](✉), Claudio Rivetta[2], Erin MacDonald[1], and David P. Chassin[2]

[1] Stanford University, Stanford, CA 94305, USA
wanlinhu@stanford.edu
[2] SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

Abstract. Human operators interact with the power control system as "in-the-loop" control elements to ensure the system stability and safety. The role of human operators becomes more critical with the increasing usage of renewable energy resources. This research seeks to support operator training by developing a technique to quantitatively model human operators' performance in the context of a simple power dispatch task. In the designed compensatory tracking task, the operator acted on the system output error and was part of a feedback control loop. The primary metric developed to evaluate and model the operator's performance is the normalized deviation, which is defined as the difference between the individual quadratic error and the averaged performance. Twenty-three human subjects participated in the experimental study. The data collected was then used to examine the proposed modeling approach and to obtain insights into possible effects of human factors. The proposed modeling technique will enable us to evaluate and compare the operator's performance to the optimal controller designed for the same task, and to design the training program that aims to align the operator's performance closer to the optimal controller.

Keywords: Human-in-the-loop · Human operator · Modeling · Power systems

1 Introduction

Human operators monitor and control power systems to make them stable and safe. They cooperate with the automatic control system and act as "in-the-loop" control elements. With the growth of renewable energy resources (e.g., solar photovoltaics, wind generation) in modern power systems, engineers seek to implement new system control and operation requirements, which inherently introduces different roles and tasks to operators. To better support human operators, we seek to understand how system variables and human factors impact operator performance. This impact should be described quantitatively such that automatic systems can interpret and respond to operators' behaviors appropriately.

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 39–54, 2019.
https://doi.org/10.1007/978-3-030-22507-0_4

Existing literature in control systems with humans has studied human actions as the input and the output. However, research incorporating modeling human behavior in feedback loops is limited, especially in the context of power systems. Moreover, empirical data from human subjects is necessary to quantitatively describe the effects of human factors such as situation awareness and knowledge on operator performance. This paper aims to (1) develop a human operator performance modeling technique in the context of simplified power dispatch system based on empirical data collected in the lab; and (2) investigate the effects of learning and the operator situation awareness on their performance. This modeling technique could enable us to model the effect of individual human behavior on the robustness and uncertainty of human-in-loop control systems, and could help improve operator performance and potentially facilitate interactions between humans and control systems.

This paper is organized as follows. Section 2 provides background on broad human-in-the-loop control systems, and a brief review of the human-in-the-loop power system practices and challenges. Section 3 presents the methodology to address the research objective, including assumptions and the identification approach of the human-in-the-loop model and the human subject study. Results and discussions are presented in Sect. 4, followed by future work and concluding statement in Sect. 5.

2 Background

In this section, we first introduce the background of human-in-the-loop control systems and example applications. We then focus on the human roles in power systems, the operator roles in modern utility control, and challenges to enhancing the performance of human-in-the-loop power system control.

2.1 Human-in-the-loop Control Systems

Automatic control can be applied to a broad range of human-machine systems. In such systems, the human acts as an active feedback control device in the control system [1], and are referred to as human-in-the-loop control systems. The automatic controller not only supports humans, but may also introduce coordination demands to human operators. Therefore, incorporating the effect of human behavior on automatic control can help with automation design and lead to improved system performance [2].

Human-in-the-loop controls have been extensively studied in aviation and automobile applications. The research focus is to enhance the stability and safety of the human-vehicle system. For example, evaluating the handling qualities is critical in piloted flight design [3]. This requires an understanding of the joint human pilot control behaviors and aircraft control system dynamics (i.e., the pilot-aircraft system dynamics) [3,4]. Employing pilot models could help us assess the handling qualities and predict pilot/aircraft performance [5]. Human-in-the-loop technology has also been used to assist the progress of the longitudinal automation system for intelligent vehicles [6] and the controller synthesis for level 3 self-driving (limited self-driving automation) vehicles [7]. Similar to the aviation applications, a human driver model could enhance the development of a safety algorithm for semi-autonomous vehicle control [8].

In addition to aviation and automobile applications, human operators are also necessary for various control systems because they are able to adapt to changing situations [9]. For example, human knowledge and decisions are currently difficult to automate in some secure systems [10]. Moreover, for applications such as assistive technology, humans must be the operator and the control system is designed to fulfill the human's demands. A wheelchair-mounted robotic arm is a concrete example for this type of human-in-the-loop system [11]. The challenge includes interfacing the human to the robotic system taking into account the limitations of the human, and reducing the human's cognitive load.

2.2 Human-in-the-loop Power Systems

Human-in-the-loop has been applied as a control element in power systems in two areas: in utility control rooms and in residences, which can be seen as operator-in-the-loop and user-in-the-loop, respectively. The latter corresponds to an indirect controller that is designed based on the preference(s) and behavior(s) of each resident. The primary objective of the energy system controller design is to reduce energy use and better align demand with intermitted supply without adversely impacting comfort. Popular examples of energy saving user-in-the-loop systems include the smart thermostat [12] and the HVAC (heating, ventilation and cooling) system in office or residential buildings [13,14].

In contrast to the role of users, human operators actively respond to and/or control power generation, transmission and distribution systems. Operators monitor the grid conditions visually in real-time using data from supervisory control and data acquisition (SCADA) systems and energy management systems (EMS) in a modern utility control facility [15]. With the advance of automation, human operators' control responsibilities have evolved toward awareness, decision-making, and response to unforeseen or contingency events. They intervene as necessary to regulate power grids and maintain system stability when automation does not function as required or expected [16]. Human interventions can have a significant positive effect on power system events [17]. However, humans can also be identified as one of the contributing causes of power grid cascading failures [18–20]. This suggests the need for a deep understanding of how to improve the effectiveness of the human operator as a preventive safety function, how we improve the interactions between human operators and existing automation systems, and how we train the human operator.

Among the various approaches to improve the human operator performance in the literature, the focus has generally been on enhancing the situation awareness and the ability of human-machine interface design to provide observability and controllability to the operator. Lackman and Söderlund recommended adjusting the automation to an optimum level such that operators are aware of the situation and are involved in decision making [17]. Aguiar et al. suggested a procedure that involves observing and analyzing the emotional component of the operator behavior during the operation of electric power systems to refine user (i.e., operator) behavior models [21]. Research in the field of human-computer interaction also inspired approaches to improve interactions between humans and machines [22,23].

The growth of renewable energy has posed new challenges to both the system designers and the operators. Despite the integration of new technologies such as artificial intelligence, machine learning, etc., to enhance the diagnostic tools available in the control room for operators, there remains a concern in the power engineering community that the higher penetration of renewable energy generation will make it more difficult for the operator to control the power system due to the larger variability and uncertainties of these new energy supply resources. Automation will increase and be applied more and more to control the system, but the operator will always be part of the control of complex loops because the ability of a human being to adapt to changing situations is almost unique [9]. Training for operators dealing with the new generation sources will be very important in the future.

3 Materials and Methods

In this section we present the initial steps taken to understand the behavior of humans in control rooms. We present a simple experimental system design that tests the degree to which a control system can identify a linear time-invariant (LTI) neuromotor response model of a human operator, separately from the mental response model that guides the human's decision-making process.

3.1 Human-in-the-loop Model—Assumptions and Model Identification

Based on the manual control cybernetic and the human controller modeling literature, the development and evaluation of control skills during training programs and the verification of the overall effectiveness and transferability of learned skills are relevant. These concepts have defined our research roadmap to investigate the design of training programs and their evaluation and quantification based on control theory. To define the framework to apply the control background in the design of the training program and its evaluation, a simple setup was created wherein the operator continuously evaluates and controls a simple power dispatch based on varying demand. This simplified setup allows both the use of LTI system models and system identification techniques to quantify the control model when the human-in-the-loop controls the system. This rigorous mathematical approach will allow the researcher to advance quickly and efficiently from the analysis stage to the design, application and evaluation of the training program.

Mathematical models of human response have been useful in predicting human operator compensation strategies and performance for a wide range of plant-operator dynamics. State-of-the-art cybernetics theory describes human controllers as (quasi-) LTI feedback systems. The most successful models applied are those which consider the human behavior in the highly-constrained compensatory tracking task, without any preview of future task constraints, allowing the operator only to react to what happens. However, the time-invariance assumption prevents us from modeling what is a defining attribute of human controllers, namely their ability to adapt to changing situations. More elaborated models and tools to identify human manual control have been presented to study and understand human learning, adaptation and the versatile set of anticipatory feed-forward control behavior. Our principal hypothesis here is that as an operator learns a compensation strategy for a new control system, the LTI model of neuromuscular system (NMS) portion of compensatory strategy converges quickly to an LTI model, even while other aspects of the system change slowly over time. This NMS learning happens relatively quickly and can be readily discerned by a suitably-design LTI system identification procedure.

A general block diagram of the human-in-the-loop controlling a dynamical plant is depicted in Fig. 1. The controller includes feedback (FB) and feed-forward (FF) blocks, to capture both the compensatory tracking and the pursuit and preview tasks, respectively. All the blocks, including the NMS, are time-variant to consider learning and adaptation in the model. This general model has the basis on the Successive Organization of Perception hierarchy for human manual control first introduced by Krendel and McRuer [24]. A recent overview on modeling of the human-in-the-loop behavior, namely *Manual Control Cybernetics*, including a large bibliography analysis has been presented in [9].

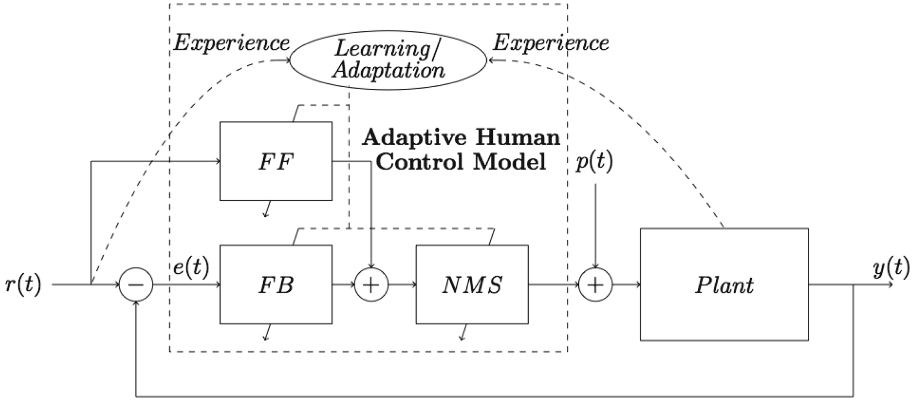


Fig. 1. System block diagram including the human controller. (Adaptive Human Control Model—FB: feedback block, compensatory task; FF: feed-forward, pursuit + preview tasks; NMS: NeuroMuscular System)

In the setup presented in this paper, the human-in-the-loop is forced to operate in the compensatory task acting solely on the error $e(t)$ between the reference $r(t)$ and the system output $y(t)$. For that, the excitation signal, acting as a perturbation $p(t)$, is created such there is not anticipatory information for the human, except that there is a short period of training to let the operator become familiar with the system mock-up implemented in the computer. The perturbing signal in the system is the power demand, which remains constant for some interval and at random times its magnitude is suddenly increased or decreased. Based on this signal the operator is only aware of a possible future change but he/she does not know when the change is going to happen nor the magnitude and sign of the demand perturbation. Additionally, the system model is not changed. Thus the operator does not need to adapt or learn new strategies to control the plant during the performance evaluation and only the NMS portion of the compensatory strategy must be learned. Based on the model in Fig. 1, only the block corresponding to the feedback (FB) should be activate following the forced compensatory tracking task.

Because the final goal of this research is to define operator training and its evaluation, the model of the human controller is defined following two paths: (i) using system identification techniques to obtain the controller models corresponding to each of the operators under evaluation, (ii) designing an optimal controller taking into account the inherent human limitations (e.g., neuromotor response, delays). In this case, the assumption is that the behavior of a highly trained and well-motivated individual will closely resemble that of an optimal controller. Based on this information the idea is to design training for the operators such that after the individuals are trained all of them will perform as close to the optimal controller as humanly possible.

If the compensatory tracking task of the individual is driven by either the reference or perturbing signal, the model of the controller is single-input-single-output and conventional identification techniques can be applied to quantify the controller block. Because the controller block is part of a feedback loop, the power spectrum of the input signal to this block is strongly affected by the sensitivity/complementary sensitivity transfer function of the closed loop system. From the identification point of view, it introduces some attenuation in the lower/higher frequency range which could compromise the identification of low/high order modes in the human response. The design of the input signal to conduct the identification has to take into account this issue, enhancing the response of the operator at low/high frequencies.

The design of the optimal controller emulating the human-in-the-loop leads to high order controllers (over-parameterized). The order of this optimal controller is equal to the sum of orders of the plant, any disturbance filter in the loop, the neuromotor lag and the equivalent Pade approximation of the delay. Applying model or controller reduction techniques directly to the full order system does not guarantee optimality of the reduced operator model. Fixed order models for the optimal response of the human operator can be obtained using optimal projection synthesis [25].

3.2 Experimental Setup and Procedure

We developed a computer-based operator workstation to allow individual participants to perform simple power dispatch tasks. The simulated workstation was designed to support system identification on the operator performing tasks. The block diagram of the controlled feedback system representing the power dispatch in a power system is shown in Fig. 2. The dynamics of the plant is mainly defined by the inertia of the rotating machines, the damping of the loads, and the system's primary frequency response, which can only arrest the frequency deviation but cannot correct it. The correction is the task left to the human operator. We note that power systems generally provide this secondary frequency control response as well, as it is relatively simple to implement. However, the purposes of this study we have constructed the experiment system such that the human must provide it, so that (a) the human task is simple and quickly understood by the test subjects, and (b) the subject's performance can be easily compared to the performance of the optimal automatic secondary frequency control response.

The system has two input signals, one is the reference $r(t)$ and the other is the perturbation signal $p(t)$. The changing power demanded by the loads is acting as a perturbation to the system and that signal is designed to stimulate the compensatory tracking task of the human controller. The reference signal is the reference frequency f_o equal to 60 Hz. In this setup, the participant ("operator") controls the generated power to balance the demanded loads and keeps the grid frequency $f(t)$ almost constant and equal to the reference frequency f_o.

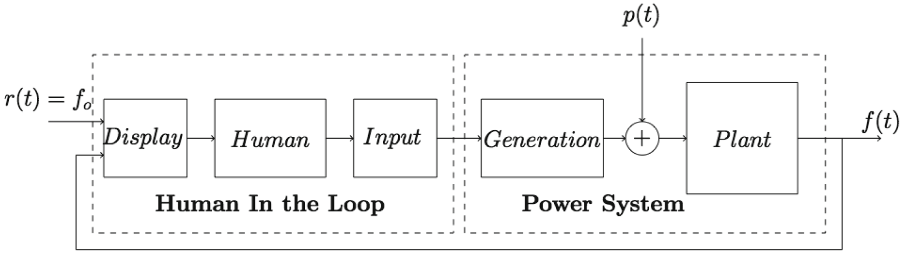


Fig. 2. Block diagram of the power system used in the experimental setup.

During the task, the demanded load is perturbed following step changes of different magnitudes at random times. The operator must increase or decrease the generator power output until the frequency reference is re-established. The screen of the workstation is depicted in Fig. 3, where the main signals and the control are displayed. The screen shows in time the load demand and the frequency error, $f(t) - f_o$ and the operator changes the generated power by acting on the arrow keys, forcing $(f(t) - f_o) \to 0$ after the perturbation at t_i.

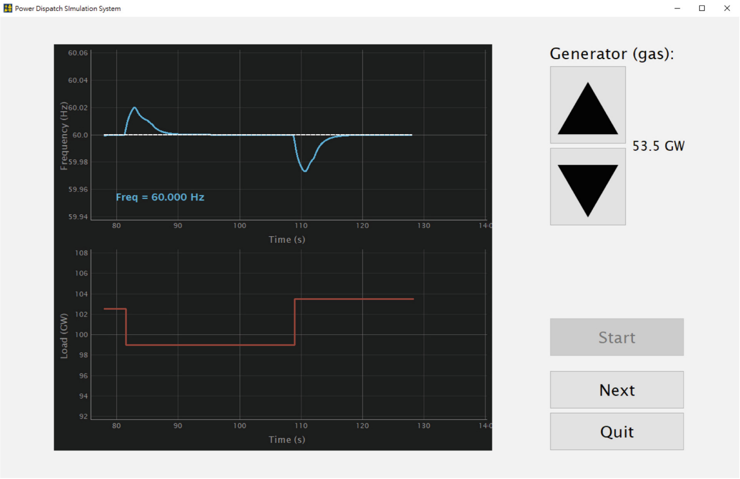


Fig. 3. The screenshot of the simulation control task.

3.3 Participants

A total of 23 individuals, age between 18 and 35 years old ($M = 22.2$ years), participated in our study. Among the participants, 9 were males and 14 were females. The participants were paid $20.00 for their participation in this study. This study was approved by the University Institutional Review Board and each participant signed the informed consent before beginning the study.

After consenting to participate in the study, participants read instructions on the screen and practiced the operation for 80 s. They read further instructions after the practice, and then performed a 12-min power dispatch task. Participants then answered a questionnaire at the end of the study.

3.4 Questionnaire

To investigate the effects of operators' background and situation awareness level on their task performance, we collected the information of their age, gender, and education level in the questionnaire. Moreover, we adapted the situation awareness rating technique (SART) developed by Taylor [26] to obtain operators' self-assessed situation awareness level. Example questions are shown in Table 1.

3.5 Operator's Performance and Human Controller Identification

Using the setup presented, Fig. 4 summarizes an operator's response to the task, where the upper plot depicts the load perturbation, the middle indicates the dispatched power (operator action), and the lower plot shows the grid frequency. These signals are used to evaluate the performance of the participants and also to identify the model of human controller.

Table 1. Example questions asked at the end of the study.

Example demographic questions
What is your age?
What is your highest degree or level of school you have completed?
Example situation awareness questions
On a scale of 1–7 how changeable is the situation?
7 - High: the situation was highly unstable and likely to change suddenly
1 - Low: the situation was very stable and straightforward
On a scale of 1–7 how many variables are changing within the situation?
7 - High: there were a large number of factors varying
1 - Low: there were very few variables changing (Low)?

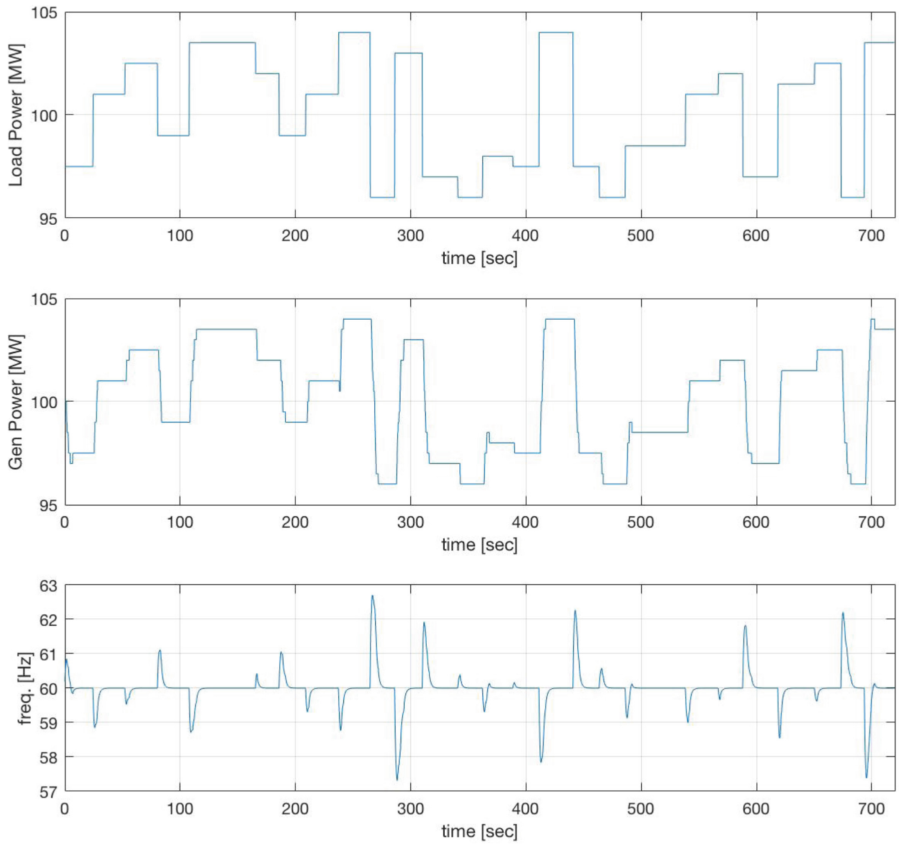


Fig. 4. The operator's response to the task. Top: load power perturbation (demand); middle: generator output; bottom: grid frequency.

One of the metrics used to evaluate the operator's performance was the quadratic error difference of the instantaneous frequency $f(t)$ and the reference value f_o equal to 60 Hz, at each period T_i after the load perturbation. It is defined as $<e^2>_i = \frac{1}{T_i}\int_{t_i}^{T_i+t_i}(f(t)-f_o)^2 dt$, where t_i denotes the i^{th} time where the perturbation is applied and $i = 1, ..., N$. In a linear system, this magnitude $<e^2>_i$ follows a quadratic relationship with respect to the perturbation magnitude p_i, $<e^2>_i = G_o^2 p_i^2$ (or $<e^2>_i^{1/2} = G_o p_i$), where G_o can be defined as an equivalent sensitivity.

When the human controller is in the feedback loop, $<e^2>_i^{1/2} = f(p_i)$ is not linear. A model to represent this function is proposed as

$$<\hat{e^2}>^{1/2} = (G_o + \mathit{\Delta} G(p))p \approx (G_o + a|p|)p \tag{1}$$

where $\mathit{\Delta} G(p) \approx a|p|$ represents the nonlinear effect of the neuromotor-actuator composite system. This term take into account the longer movement or action on the actuator that the operator has to execute to compensate increasing perturbations (magnitude). Using this approximation, the data is fitted to the model defining the overall average performance of the operator. Figure 5 shows an example analysis of one participant. From this curve, it is possible to define two metrics, G_o and a, to compare and quantify the performance of different operators.

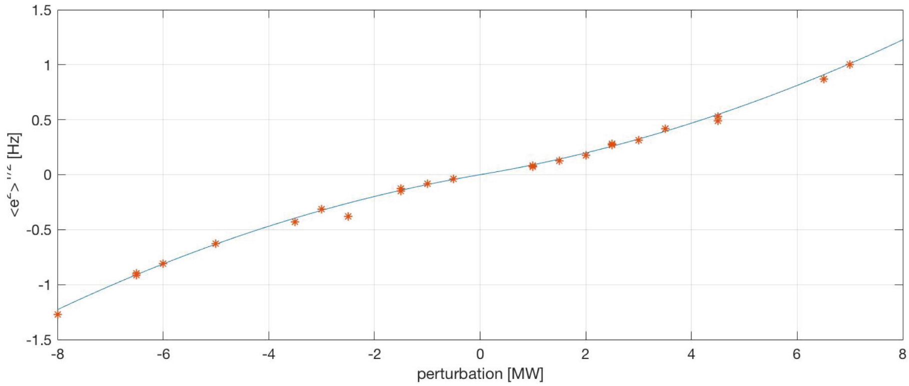


Fig. 5. Example performance analysis of one participant. The quadratic error $<e^2>^{1/2}$ is a function of the perturbation magnitude.

Another metric to define the operator's performance from this analysis is the normalized deviation. It is defined as the difference between the individual quadratic error $<e^2>_i^{1/2}$, calculated during the interval $[t_i, t_i + T_i)$, with respect to the averaged curve $<\hat{e^2}>^{1/2} = (G_o + \mathit{\Delta} G(p))p$. It is,

$$dev_i = \frac{<e^2>_i^{1/2} - <\hat{e^2}>_i^{1/2}}{p_i}. \tag{2}$$

This metric can be used to trace how the operator's performance evolves with time during the task. It allows us to study the effect of learning, concentration/fatigue, etc. Figure 6 shows the operator's performance at each load perturbation.

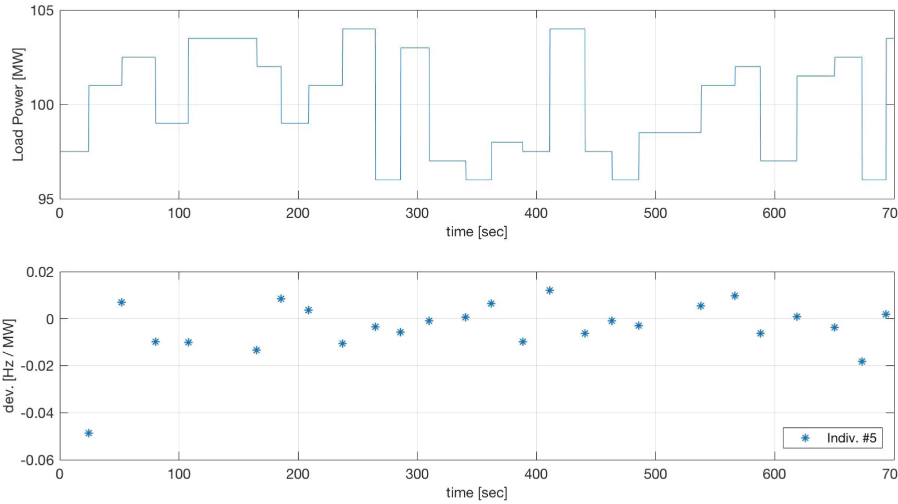


Fig. 6. An operator's learning performance (defined as the normalized deviation with respect to the average performance) at each load perturbation.

These results can be compared with the outcome of the system analyzed in simulation (or a real system with the human controller replaced by a hardware controller). In that case, the quadratic error $<e^2>_i$ after each perturbation will follow without dispersion the average curve $<e^2>^{1/2} = G_o\, p$, achieving the minimum G_o for the case of the optimal controller design.

As noted above, this system is driven by two inputs: the reference $r(t) = f_o$, that is the reference frequency and it is kept constant and the perturbation $p(t)$, that is used in this case, to inject an exogenous signal to drive the operator in compensatory tracking task. It is important to remark that the operator observes the error $e(t)$ between the reference frequency f_o and the grid frequency $f(t)$ and try to keep this error close to zero when the load demand signal $p(t)$ perturbs the system. The feedback loop introduces some filtering to the perturbing signal $p(t)$ such that the spectral content of the error signal $e(t) = f_o - f(t)$ (Fig. 1), used to identify the human controller, is attenuated at low and high frequencies. The signal used as perturbation to excite the system has some features to enhance the reaction of the operator at low and high frequencies and also allows us to characterize completely his/her behavior as a controller at each instant that the perturbation is applied. With this signal design, a family of transfer functions for the operator as the equivalent controller can be calculated. This family of transfer functions allows us to mathematically compare the operator's behavior with an optimal operator or controller for such a system.

Based on the collected data, the relationship in frequency domain between the frequency error, $e(t) = f_o - f(t)$, (stimulus) and the operator's action defines the transfer function of the human controller. Assuming a linear time-invariant model, the controller can be represented by an ARMA filter:

$$C(s) = \frac{\sum_{i=0}^{m} b_i s^i}{\sum_{i=0}^{n} a_i s^i}, \tag{3}$$

with the order $n \geq m$, the complex frequency $s = \sigma + j\omega$ and coefficients a_i, $b_i \in \mathbf{R}$ to be defined via the identification. The order of the ARMA model n is defined to capture the higher order modes of the operator's response. This transfer function represents mathematically the operator's behavior at any power change.

4 Results and Discussions

The results obtained from the early trials show the promise of characterizing a subject's neuromotor lag compensation learning rate, as well as the potential emergence of fatigue or boredom. Figure 7 shows operator's performance with respect to the average performance (i.e., the fitting curve) of the operator for different individuals. For each individual, the data was fitted to the average curve (1) and the parameter G_o characterizing the performance of each operator spanned between $G_o \in [0.0691, 0.195]$ Hz/MW. The normalized deviation was between +0.249/ −0.188 Hz/MW for all participants.

From those particular plots, it is possible to observe the difference in performance between individuals. For example, individuals #15, 16, 19 and 5 (Fig. 6) corresponded to both the minimum deviation with respect to the average performance of the operator and the best performance ($G_o \in [0.0681, 0.085]$ Hz/MW). Individual #15 showed a learning process at the beginning of the task, and after that he/she performed better than individuals #5, 16 and 19. On the other hand, Indiv. #3 performed well at the beginning, but showed a sign of losing concentration or feeling fatigue and experienced degrading performance toward the end of the test.

The quantification of the operator's response using the identified transfer function (Eq. 3) is summarized in Fig. 8. In this figure, the transfer function is calculated at each load power transition, defining a family of 'controllers' that characterize the unique response of the operator to each perturbation. This family of transfer functions gives a quantification of the deviation of the response of the operator to the same task with different strength.

To evaluate whether or not the self-reported situation awareness level has an effect on the operator's performance, we conducted two-sample t-tests with the performance as the dependent variable and the situation awareness reported on each SART question as the independent variables. Results suggest that two aspects of the situation awareness had a significant effect on the performance.

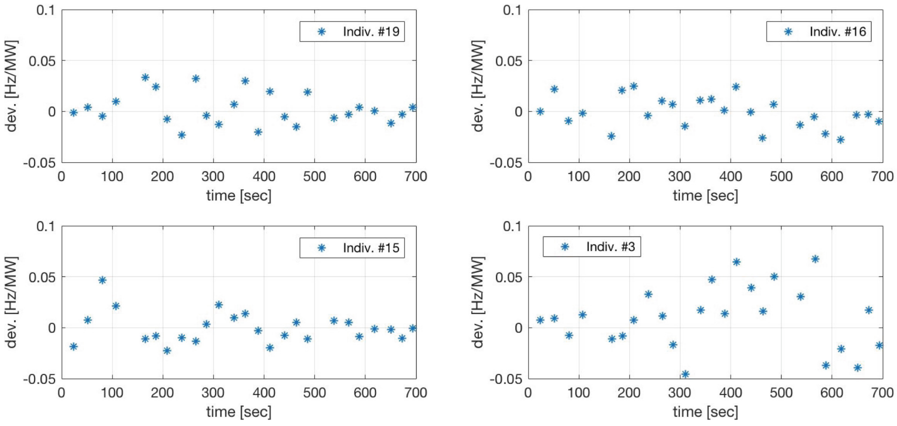


Fig. 7. The operator's performance with respect to the average performance of the operator for different individuals.

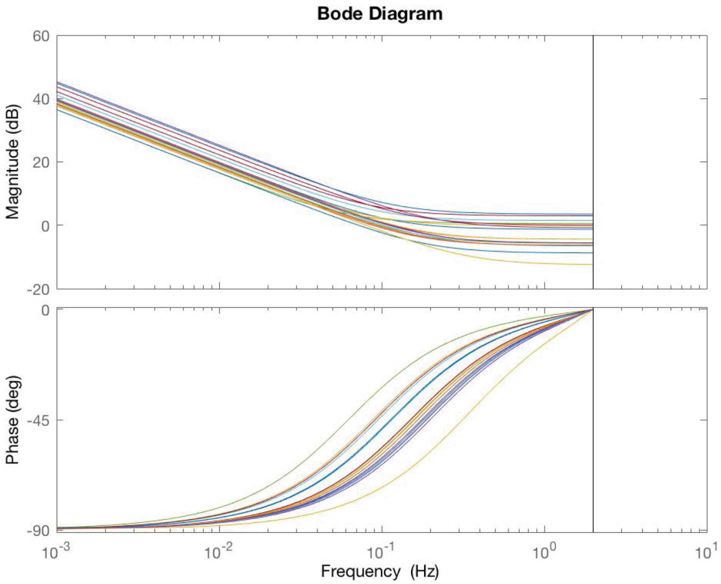


Fig. 8. Frequency response of the human controller (Indiv. #1) to each power perturbation.

One factor was *how much was the operator's attention divided in the situation* (p-value = 0.010) and the other was *how familiar was the operator with the situation during the task* (p-value = 0.003). To be specific, individuals who were focused on few aspects of the situation performed better (i.e., smaller error $<e^2>_i$) than those who were concentrating on many aspects. Our simulation control task provides frequency plot, the number of the current frequency, the system load plot, and the current generator output to support the operation. These can be seen as inputs to the human controller. Focusing on fewer aspects of the situation could imply a higher attention to the relationship between a single input and the frequency deviation. With the simplified nature of the simulation task, this behavior may result a fast response and thus reduce the error. In addition, individuals who perceived the situation as new performed better than those who had a great deal of relevant experience. This is different from what we expected since the literature reported a positive effect of experience on the performance. Nonetheless, perceiving the situation as new may lead to a cautious action and better performance. Future work should involve evaluating these effects with a complex task, and further defining the experience in a given context.

5 Closing Remarks and Future Work

This research developed a technique to quantitatively model human operators' performance in a simple power dispatch task by characterizing the human as part of a feedback system. With the empirical data collected from human subjects, we can observe the effects of some human factors such as learning/practice from individual operators. Moreover, between-subjects comparison enabled use to identify the effects of situation awareness on the performance. Furthermore, the developed modeling technique is based on control theory, which enables us to: (1) compare the response of the operator with the optimal controller designed for the same plant or task, (2) define training for the operators to translate their performance closer to the optimal controller [25] and reduce the deviation of the response from one perturbation to the next.

The next step in our research is twofold. Firstly, we will investigate with this setup the impact of the reaction delay of the individual and how the operator adapts to minimize the error. In parallel, we will investigate mental decision-making models for system operators who are trained to follow "play-books" for responding to system contigencies. We will incorporate observations of how system operators work with more complex systems, using the identified neuromotor compensation models to distinguish mental decision-making delays from perception and actuation delays.

Acknowledgement. This work was supported by Laboratory Directed Research and Development funding at SLAC National Accelerator Laboratory, which is operated for the US Department of Energy by Stanford University under Contract No. DE-AC02-76SF00515.

References

1. Hess, R.A.: Human-in-the-loop control. In: Levine, W.S. (ed.) Control System Applications, pp. 327–335. CRC Press, Boca Raton (1999)
2. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man. Cybern. - Part A: Syst. Hum. **30**(3), 286–296 (2000)
3. Harper, R.P., Cooper, G.E.: Handling qualities and pilot evaluation. J. Guidance Control Dyn. **9**(5), 515–529 (1986)
4. Damveld, H.J., Van Paassen, M.M., Mulder, M.: Cybernetic approach to assess aircraft handling qualities. J. Guidance Control Dyn. **34**(6), 1886–1898 (2011)
5. Hess, R.A.: Analytical assessment of performance, handling qualities, and added dynamics in rotorcraft flight control. IEEE Trans. Syst. Man Cybern. - Part A: Syst. Hum. **39**(1), 262–271 (2009)
6. Chiang, H.H., Wu, S.J., Perng, J.W., Wu, B.F., Lee, T.T.: The human-in-the-loop design approach to the longitudinal automation system for an intelligent vehicle. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **40**(4), 708–720 (2010)
7. Li, W., Sadigh, D., Sastry, S.S., Seshia, S.A.: Synthesis for human-in-the-loop control systems. In: Ábrahám, E., Havelund, K. (eds.) TACAS 2014. LNCS, vol. 8413, pp. 470–484. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54862-8_40
8. Driggs-Campbell, K., Shia, V., Bajcsy, R.: Improved driver modeling for human-in-the-loop vehicular control. In: 2015 IEEE International Conference on Robotics and Automation, pp. 1654–1661 (2015)
9. Mulder, M., et al.: Manual control cybernetics: state-of-the-art and current trends. IEEE Trans. Hum.-Mach. Syst. **48**(5), 468–485 (2018)
10. Cranor, L.F.: A framework for reasoning about the human in the loop. In: Proceedings of the 1st Conference on Usability, Psychology, and Security (2008)
11. Tsui, K.M., Behal, A., Kontak, D., Yanco, H.A.: "I want that": human-in-the-loop control of a wheelchair-mounted robotic arm. Appl. Bionics Biomech. **8**, 127–147 (2011)
12. Lu, J., et al.: The smart thermostat: using occupancy sensors to save energy in homes. In: Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems - SenSys 2010, New York, USA. ACM Press, New York (2010)
13. Zeiler, W., Houten, R., Boxem, G., Vissers, D., Maaijen, R.: Indoor air quality and thermal comfort strategies: the human-in-the-loop approach. In: Proceedings of International Conference on Enhanced Building Operations (2011)
14. Nakaya, T., Matsubara, N., Kurazumi, Y.: Use of occupant behaviour to control the indoor climate in Japanese residences. In: Proceedings of Conference: Air Conditioning and the Low Carbon Cooling Challenge, Windsor, UK, pp. 27–29 (2008)
15. Giri, J., Parashar, M., Trehern, J., Madani, V.: The situation room: control center analytics for enhanced situational awareness. IEEE Power Energy Mag. **10**, 24–39 (2012)
16. Obradovich, J.H.: Understanding cognitive and collaborative work: observations in an electric transmission operations control center. In: Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting, pp. 247–251 (2011)
17. Lackman, T.O., Söderlund, K.: Situations saved by the human operator when automation failed. Chem. Eng. Trans. **31**, 385–390 (2013)
18. Horowitz, B.S.H., Phadke, A.G., Renz, B.A., Horowitz, S., Phadke, A.G.: The future of power transmission. IEEE Power Energy Mag. **8**, 34–40 (2010)

19. Brummitt, C.D., Hines, P.D.H., Dobson, I., Moore, C., D'Souza, R.M.: Transdisciplinary electric power grid science. Proc. Nat. Acad. Sci. **110**(30), 12159 (2013)
20. Hines, P., Balasubramaniam, K., Sanchez, E.C.: Cascading failures in power grids. IEEE Potentials **28**, 24–30 (2009)
21. Aguiar, Y., Vieira, M., Galy, E.: Refining a user behaviour model based on the observation of emotional states. In: COGNITIVE 2011: The Third International Conference on Advanced Cognitive Technologies and Applications Refining, pp. 36–40 (2011)
22. Dumas, J.S., Redish, J.: A Practical Guide to Usability Testing. Intellect Books, Portland (1999)
23. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 249–256. ACM (1990)
24. Krendel, E.O., McRuer, D.: A servomechanism approach to skill development. J. Franklin Inst. **269**(1), 24–42 (1960)
25. Doman, D.B., Anderson, M.R.: A fixed-order optimal control model of human operator response. Automatica **36**, 409–418 (2000)
26. Taylor, R.: Development of the situational awareness rating technique (sart) as a tool for aircrew systems design. In: AGARD Conference Proceedings, CP478, Copenhagen, DK (1990)

Comparing Eye-Gaze Metrics of Mental Workload in Monitoring Process Plants

Wenyan Huang[1], Yunshu Xu[1], Michael Hildebrandt[2],
and Nathan Lau[1](✉)

[1] Virginia Tech, Blacksburg, USA
{whuang, yunshuxu, Nathan.lau}@vt.edu
[2] Institute for Energy Technology (IFE), Halden, Norway
michael.hildebrandt@ife.no

Abstract. Continuous, non-invasive workload indicators of operators is an essential component for dynamically assigning appropriate amount of tasks between automation and operators to prevent overload and out-of-the-loop problems in computer-based procedures. This article examines the monitoring task difficulty manipulated by the task type and load, and explores physiological measurements in relation to mental workload. In a within-subject design experiment, forty-five university students performed monitoring tasks in simulated nuclear power plants (NPPs) control room. The performance of monitoring tasks (accuracy), subjective mental workload (NASA Task Load Index), as well as four eye-related physiological indices were measured and analyzed. The results show that as monitoring task difficulty increased, task performance significantly decreased while NASA-TLX, number of fixations, and dwell time significantly increased. Number of fixations and dwell time could be effective, non-invasive continuous indicators of workload for enabling adaptive computer-based procedures.

Keywords: Workload · Eye-related measures · NASA-TLX · Process plants

1 Introduction

Power generation and petrochemical plants rely on procedures extensively [1, 2]. Traditionally, procedures were written on paper, and remain so in many plants. However, there are significant limitations of voluminous paper procedures due to complexity and mental demand with equipment and operations [2–6]. For example, Kontogiannis [3] concluded that paper procedures were inadequate in presenting complex instructions, handling cross-references, tracing suspended or incomplete steps, and monitoring procedural progress. Ockerman and Pritchett [2] also found that paper procedures can be too heavy, delicate, immobile and difficult to follow, preventing operators from executing procedures efficiently.

Computerized procedure systems (CPs) are being developed to resolve the limitations of paper procedures [3, 4, 7–11]. CPs are digital versions of paper procedures that may include additional capabilities to support the operators in executing procedures. These capabilities range from hyperlinks connecting different parts of a

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 55–65, 2019.
https://doi.org/10.1007/978-3-030-22507-0_5

procedure, dynamic displays presenting parameters or controls relevant to procedural steps being executed, automatic checking of preconditions, or automatic execution of control commands [12–14]. CPs can aid process plant operators in reducing operation time and errors while alleviating overall workload. For example, Huang and Hwang [7] showed that average operation time and errors for executing decision and action tasks to deal with alarm signals were significantly reduced with CPs compared to paper procedures.

The benefits of CPs may come with the risk of out-of-the-loop (OOTL) performance problems, the decreased ability of the human operator to intervene or assume manual control when automation fails [10, 15, 16]. Specifically, relieving operators from manually checking pre-conditions and executing control actions to reduce workload may lead them to lose track of procedural steps and misjudge plant state if they haphazardly accept any recommendation of the CPs [12]. Consequently, operators may not abort an inappropriate procedure when the CPs are incorrect, or take inappropriate actions due to wrongly assumed plant state. Taking an inappropriate action might include the control room operator calling field operators to fix equipment that is in an unsafe state because the CPs have changed the equipment setting without operator awareness. For example, when a return-to-normal alarm is reset automatically by CPs, operators may not be aware that such an alarm had sounded, hindering the operator's comprehension and prediction of the plant state [17].

Adaptive automation [18–21] has been proposed as a solution for balancing risk of experiencing OOTL and workload problems. Specifically, real-time assessment of workload can be used to determine appropriate amount of tasks, thereby keeping operators engaged and preventing the OOTL problem [19, 22]. Thus, CPs adaptive to operator workload on monitoring and controlling process plants may reduce the risk of excessive workload and OOTL problem.

As the first step towards developing CPs adaptive to operator workload, this study investigated the use of eye-gaze metrics for assessing operators' workload in monitoring process plants. Specifically, we examined the relationships between eye-gaze measures with respect to a subjective rating scale of workload and task performance. Further, we examined which eye-gaze measures would be most sensitive to manipulation of task difficulty that impacts workload.

1.1 Continuous Indicator of Workload with Eye-Tracking

Eye-tracking can provide nonintrusive, continuous indicators of mental workload experienced by process plant operators, whose tasks involve substantial visual (monitoring) and cognitive processing (diagnosis and self-regulation) [23]. Eye movements are motor responses that are regulated by the cortical and subcortical brain system [24], providing information on the distribution of attention in terms of what stimuli are attended to, for how long, and in what order [25]. Substantial research indicates a correlation between human cognitive workload and eye activity measures, including fixations, saccades and blinks [26–32].

Lin et al. [33] argued that eye fixation and pupil diameter parameters are sensitive indicators to access mental workload. New information is mainly acquired during fixations [24, 34], as suggested by the eye-mind hypothesis postulating that what is being fixated by the eyes indicates what is being processed in the mind [35]. Only under limited special circumstances can new information be acquired during saccades [36, 37]. Larger number of fixations implies a large magnitude of required information processing and hence higher workload. Longer fixation duration suggests more time spent on interpreting, processing or associating a target with its internalized representation and thus higher workload [33, 38]. Marquart et al. [30] reviewed and concluded that dwell time, the period for a contiguous series of one or more fixations within an area of interest (AOI), can be an indicator of mental workload. Dwell time tends to increase with increasing mental task demands. Pupil diameter usually increases in response to increased difficulty levels of tasks translating to another common indicator of mental workload [6, 26, 39, 40].

1.2 Overview of This Study

Empirical research on eye-tracking for workload assessment in process control appears insufficient for developing adaptive CPs. For this reason, we conducted an experiment involving human participants performing monitoring tasks to provide further empirical evidence on whether eye-gaze measures can be effective, continuous workload indicators.

For monitoring process plants, we hypothesize that eye-gaze measures would be able to reveal the types of monitoring tasks imposing different workload. Also, these eye-gaze measures would reveal difference in task load for the same type of tasks. We further hypothesize that the NASA TLX, a validated subjective workload measure [41, 42], would correlate positively with number of fixations, average fixation duration, dwell time and pupil diameter.

2 Method

2.1 Participants

This experiment recruited 45 Virginia Tech graduate and undergraduate students (age range 18–26, 29 females and 28 males). All participants had normal or corrected-to-normal vision. Participants were compensated $10/h for about 1.5 h of their time.

2.2 Experimental Apparatus

The experiment was conducted in a quiet room, with a computer workstation presenting an overview display of a nuclear power plant on a 24″ LED monitor with 1920 × 1200 resolution at 60 Hz. Further, the computer workstation collected eye-gaze and heart rate data with the following equipment:

1. SensoMotoric Instruments (SMI) Remote Eye-tracking Device (REDn) recorded eye-gaze data at 60 Hz sampling rate. The REDn sensor was physically attached to the bottom of the monitor and connected to the computer workstation that had the SMI iVIEW software installed for data collection.

2. Shimmer3 ECG Sensors (ECG) recorded electrocardiogram (ECG), the pathway of electrical impulses through the heart muscle, sampling at 1000 Hz. The ECG was wirelessly connected to the computer workstation through Bluetooth. ECG analysis is beyond the scope of this paper.

2.3 Experimental Manipulation

The participant tasks were to identify parameters deemed out-of-range on an overview display of a fictional nuclear power plant (Fig. 1). The contents of the overview display consisted of tanks, pumps, heat exchangers and valves associated with various process parameters such as level (i.e., %), flow rate (i.e., gpm), temperature (i.e., °C), and pressure (i.e., psig & KPph), The locations of these process parameters were the same for all trials but the values of these process parameters were updated for each trial. The Question Box prompted the participants to complete two types of monitoring tasks.

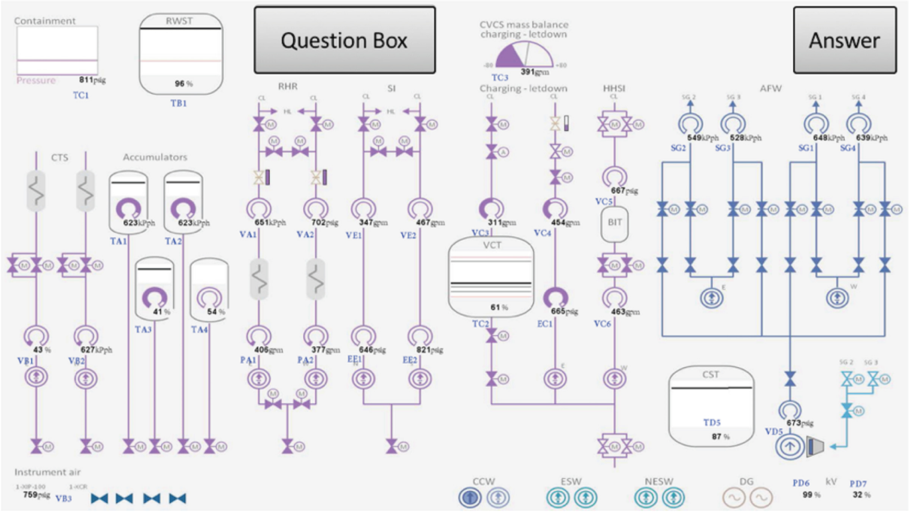
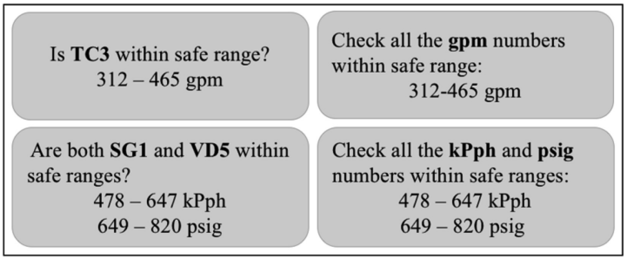


Fig. 1. Nuclear control room monitoring simulation platform presenting the overview display and monitoring tasks.

The two types of monitoring task were target-driven and series-driven verification of process parameters to represent common activities specified by procedures of industrial plants.

Target-driven verification (Fig. 2 left column). Participants were instructed to check specific targets (e.g. TC3, SG1 or VD5) per question or monitoring task. The target-driven task included either one or two targets of parameters to represent low and high task load, respectively.

Series-driven verification (see Fig. 2 right column). Participants were instructed to check all values for a specific type of parameters (e.g. gpm, kPph, psig or %) per



Is **TC3** within safe range?
312 – 465 gpm

Check all the **gpm** numbers within safe range:
312-465 gpm

Are both **SG1** and **VD5** within safe ranges?
478 – 647 kPph
649 – 820 psig

Check all the **kPph** and **psig** numbers within safe ranges:
478 – 647 kPph
649 – 820 psig

Fig. 2. Examples of question boxes on nuclear control room monitoring simulation, demonstrating manipulation of monitoring task types and task loads.

question. The series-driven task included either one or two series of parameters to represent low and high task load, respectively.

2.4 Procedure

Participants were welcomed with a brief introduction about the study in front of the computer workstation. Then they were asked to give consent and complete a health history questionnaire. The experimenter provided instructions of the control room monitoring task and answered participants' questions.

The participants completed four blocks of control room monitoring tasks for all combination of task type and load conditions: two task types (i.e., target-driven vs series driven) and two task loads (i.e., low vs. high). At the beginning of each block, the participants first completed REDn 9-point eye-gaze calibration. Participants completed a NASA-TLX questionnaire at the end of each four blocks. For each trail of the monitoring task, participants responded by clicking the corresponding out-of-range parameter(s) on the display with a mouse and then clicked the 'Answer' button (see Fig. 1) to submit their responses and proceed to next trial. The experimenter stopped REDn recording at the end of each block.

After completing four blocks of trials, the experimenter helped participants to take off the physiological instruments. Participants were given the opportunity to ask any further questions and $15 for compensation at departure.

2.5 Experimental Design

The experiment was a 2 × 2 within-subjects design with two treatments: (1) task type (singular targets or series of targets) and (2) task load (low and high). Four blocks of control room monitoring tasks were assigned in a random order across participants. Each block consisted of 3 min of monitoring tasks. Participants performed tasks at their own pace, leading to different numbers of completed trials in one block.

2.6 Measures

Participants were assessed on three categories of measures: task performance, NASA-TLX, and eye-related measures.

Response Accuracy. The response accuracy was used to assess the task performance. This measure was defined as the percentage of trials for which participant submitted the correct answer by identifying all the out-of-range parameters.

TLX Total Score. The NASA-TLX questionnaire was used to assess the subjective ratings of workload, using a 10-point visual analog scale. This questionnaire is a multidimensional instrument that consists of 6 subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. The TLX total score was computed by a combination of the six dimensions, resulting in an overall workload scale between 0 and 60.

Eye-Gaze Measures. Number of fixations, fixation duration, dwell time and pupil diameter were used as continuous indicators of workload. Area of interest (AOI) was defined as display area covering the graphic and numerical reading of the parameter(s) that should be monitored in each trial. The AOIs varied between trials depending on the monitoring task type and load. For example, a square was marked as the AOI for the trials with the one-target driven verification task, while eight squares were marked as the AOI for trials with one of the series driven monitoring task. Fixation-based metrics on AOI were extracted to indicate workload. All eye-gaze metrics were computed with SMI BeGaze software. Four metrics were selected for comparison: the total number of fixations on AOIs, average duration of a fixation on AOIs, dwell time (total fixation durations on AOIs), and pupil diameter for fixations on AOIs.

3 Results

The experiment yielded 180 observations (45 participants x 4 experimental blocks), of which twelve were removed due to participants performing the monitoring tasks incorrectly. We further failed to collect NASA TLX for an additional participant. Thus, except for NASA TLX, Pearson-product moment correlation statistics were computed to examine relationships between measures across the 168 observations and two-way analysis of variance (ANOVA) were conducted to examine differences between four experimental conditions. Statistics associated with NASA TLX only contains 167 observations.

Response accuracy was correlated with number of fixations ($r = -0.313$; $p < 0.001$) and dwell time ($r = -0.265$; $p < 0.001$). However, only pupil diameter significantly correlated with NASA TLX ($r = -0.186$; $p < 0.05$). Between eye-gaze measures, dwell time significantly correlated with all three other eye-gaze measures, including number of fixations ($r = 0.877$; $p < 0.001$), fixation duration (0.458; $p < 0.001$), and pupil diameter ($r = 0.173$; $p < 0.05$) (Table 1).

Experimental effects on response accuracy and TLX total score were examined with the nonparametric Kruskal-Wallis rank sum test because their error residuals were not normally distributed.

The nonparametric test results confirmed the hypotheses in revealing that series-driven monitoring tasks significantly hindered response accuracy ($\chi^2(1) = 31.864$, $p < 0.001$, $N = 168$). Further, the nonparametric test also revealed that task load marginally decreased response accuracy ($\chi^2(1) = 2.854$, $p = 0.091$, $N = 168$) and

Table 1. Correlation matrix of the five measures

	Response accuracy	TLX total score	Number of fixations	Avg. fixation duration	Dwell time	Pupil diameter
Response accuracy	1	-0.043	-0.313***	0.034	-0.265***	-0.003
TLX total score	-0.043	1	0.012	-0.152	-0.052	0.186*
Number of fixations	-0.313***	0.012	1	0.056	0.877***	0.068
Avg. fixation duration	0.034	-0.152	0.056	1	0.458***	0.290
Dwell time	-0.265***	-0.052	0.877***	0.458***	1	0.173*
Pupil diameter	-0.003	0.186*	0.068	0.290	0.173*	1

*** = p<0.001; ** = p<0.01; * = p < 0.05

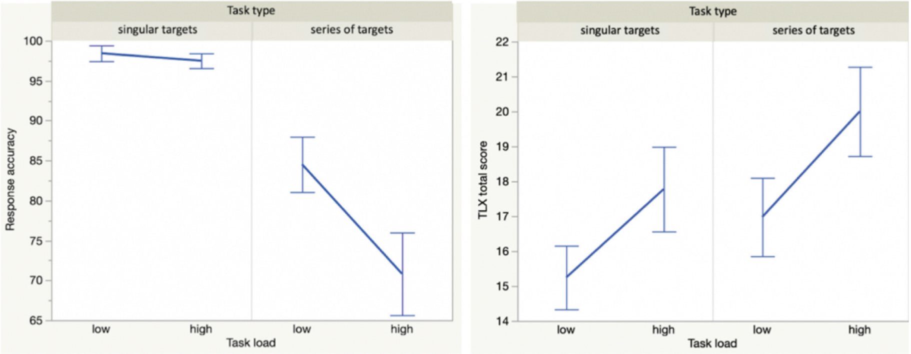


Fig. 3. Mean and standard error plots of response accuracy (left) and subjective workload rating score (right) for each combination of task type and task load.

significantly increased subjective workload ($\chi^2(1) = 4.748$, p = 0.029, N = 167) (Fig. 3).

All eye-related measures were analyzed in two-way ANOVAs. The main effect of task type was also significant on the number of fixations (F(1, 159) = 110.634, p < 0.001) and dwell time (F(1, 159) = 49.117, p < 0.001). Similarly, the main effect of task load was significant on both number of fixations (F(1, 159) = 31.5963, p < 0.001) and dwell time (F(1, 159) = 14.320, p < 0.001). Furthermore, the interaction effect of task type and load was significant on both number of fixations (F(1, 159) = 11.997, p < 0.001) and dwell time (F(1, 159) = 6.162, p = 0.014). In other words, increased task load had significantly more impact for performing series-driven than target driven monitoring tasks. However, average duration per fixation and pupil diameter did not reveal any significant effect (Fig. 4).

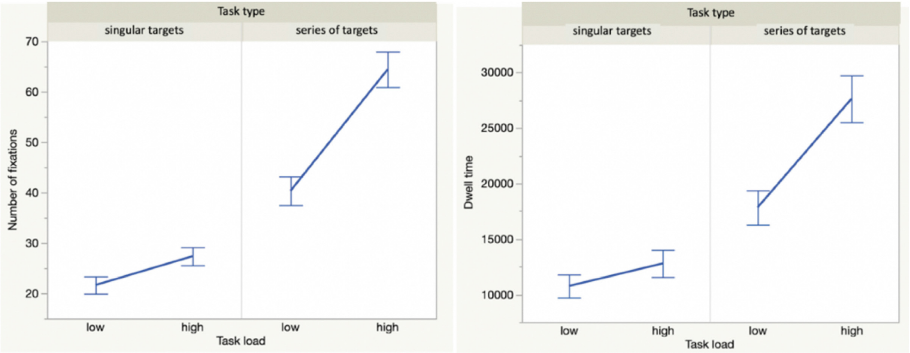


Fig. 4. Mean and standard error plot of the number of fixations (left) and dwell time (right) for each combination of task type and task load.

4 Discussion

The significant main effect of task type and load on response accuracy and NASA TLX confirmed our hypotheses, indicating that the two experimental manipulations were effective at manipulating workload. Thus, we can confidently interpret the eye-gaze metrics with respect to the response accuracy and NASA TLX measures. The number of fixations on AOIs and dwell time on AOIs showed the same main effects as response accuracy and NASA TLX, indicating the sensitivity of these two eye-gaze measures to experimental manipulations. However, these two measures were not sensitive to subjective workload because they did not correlate with NASA TLX. Pupil diameter failed to reveal any significant effects but correlated with NASA TLX. In other words, pupil diameter was sensitive to subjective workload but not to the effect of task type and load. The average fixation duration per AOI did not appear to be a sensitive measure, failing to reveal any significant correlations and experimental effects.

The results of this experiment illustrate how careful consideration is needed in selecting eye-gaze metrics for indicating workload in monitoring process plants. None of the eye-gaze measures showed significance to both correlation with NASA TLX and experimental manipulations (i.e., task type and load), so there is no clear contender of an eye-tracking measure for indicating workload. (Dwell time and number of fixations only showed significant correlation with response accuracy.)

These eye-gaze results must be also be interpreted with respect to the monitoring tasks designed for this experiment. Specifically, there are more targets for the series-driven than target-driven task type, and for the high than low task load. Thus, the number of fixations may be higher inherently due to the task characteristic of more targets rather than higher mental workload. For this reason, dwell time on AOIs might be a more robust indicator than number of fixations because dwell time is bounded by the allotted time for the block (i.e., 3 min). In the context of this study, the issue on number of targets probably does not present a significant problem for two reasons. First, having more targets is intrinsically linked to the demand of the monitoring tasks, so the results should still be representative for monitoring process plants. Second, dwell time revealed the same experimental effects as NASA TLX, lending empirical support that the experimental manipulations affect dwell time and mental workload similarly.

Another notable result is the weak, positive correlation between dwell time and response accuracy, indicating that eye-gaze behaviors could contribute to task performance. Thus, dwell time might also offer modest and continuous indication of operator engagement with system operations.

The overall empirical results indicated that dwell time could be an effective alternative to NASA TLX as a workload indicator in the context of monitoring parameters prescribed by procedures. Dwell time can be collected in a less invasive manner than NASA TLX while providing continuous indication of workload and engagement. That is, NASA TLX requires interruption of the work tasks whereas the remote eye-tracker can continuously estimate dwell time without any interference. NASA TLX might simply reflect operator initial and final impression of the given task as opposed to their level of cognitive processing as they perform the given task. Once again, the generalization of the study results is limited to task load driven by number of targets.

This research represents the early effort to integrate the concept of adaptive automation into CPs. The results of this experiment highlight the potential of various eye-gaze measures as a continuous indicator of workload to support adaptive features in CPs for control room operators monitoring process plants. Valid and reliable eye-gaze metrics of workload can support continuous, unobtrusive assessment of workload as well as adaptive aiding for display design in the main control room. Future work can examine the use of regression-based machine learning methods on multiple eye-gaze measures to indicate workload while monitoring process plants (see [43]).

References

1. Ludwig, E.E.: Applied Process Design for Chemical and Petrochemical Plants, vol. 2. Gulf Professional Publishing, Houston (1997)
2. Ockerman, J., Pritchett, A.: A review and reappraisal of task guidance: aiding workers in procedure following. Int. J. Cogn. Ergon. **4**(3), 191–212 (2000)
3. Kontogiannis, T.: Applying information technology to the presentation of emergency operating procedures: implications for usability criteria. Behav. Inf. Technol. **18**(4), 261–276 (1999)
4. Niwa, Y., Hollnagel, E., Green, M.: Guidelines for computerized presentation of emergency operating procedures. Nucl. Eng. Des. **167**(2), 113–127 (1996)
5. Park, J., Jung, W.: The operators' non-compliance behavior to conduct emergency operating procedures—comparing with the work experience and the complexity of procedural steps. Reliab. Eng. Syst. Saf. **82**(2), 115–131 (2003)

6. Xu, S., et al.: An ergonomics study of computerized emergency operating procedures: presentation style, task complexity, and training level. Reliab. Eng. Syst. Saf. **93**(10), 1500–1511 (2008)
7. Huang, F.H., Hwang, S.L.: Experimental studies of computerized procedures and team size in nuclear power plant operations. Nucl. Eng. Des. **239**(2), 373–380 (2009)
8. Hwang, F.H., Hwang, S.L.: Design and evaluation of computerized operating procedures in nuclear power plants. Ergonomics **46**(1–3), 271–284 (2003)
9. Landry, S., Jacko, J.: Improving pilot procedure following using displays of procedure context. Int. J. Appl. Aviat. Stud. **6**(1), 47–70 (2006)
10. Lee, S.J., Seong, P.H.: Development of an integrated decision support system to aid cognitive activities of operators. Nucl. Eng. Technol. **39**(6), 703 (2007)
11. Lin, C.J., Hsieh, T.L., Yang, C.W., Huang, R.J.: The impact of computer-based procedures on team performance, communication, and situation awareness. Int. J. Ind. Ergon. **51**, 21–29 (2016)
12. Naser, J.: Computerized procedures: design and implementation guidance for procedures, associated automation and soft controls, vol. 1015313, Draft Report. EPRI (2007)
13. Yang, C.W., Yang, L.C., Cheng, T.C., Jou, Y.T., Chiou, S.W.: Assessing mental workload and situation awareness in the evaluation of computerized procedures in the main control room. Nucl. Eng. Des. **250**, 713–719 (2012)
14. Fink, R.T., Killian, C.D., Hanes, L.F., Naser, J.A.: Guidelines for the design and implementation of computerized procedures. Nucl. News **52**(3), 85 (2009)
15. Kaber, D.B., Endsley, M.R.: The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. Theor. Issues Ergon. Sci. **5**(2), 113–153 (2004)
16. Lin, C.J., Yenn, T.C., Yang, C.W.: Automation design in advanced control rooms of the modernized nuclear power plants. Saf. Sci. **48**(1), 63–71 (2010)
17. Huang, F.H., et al.: Experimental evaluation of human–system interaction on alarm design. Nucl. Eng. Des. **237**(3), 308–315 (2007)
18. Byrne, E.A., Parasuraman, R.: Psychophysiology and adaptive automation. Biol. Psychol. **42**(3), 249–268 (1996)
19. Kaber, D.B., Riley, J.M.: Adaptive automation of a dynamic control task based on secondary task workload measurement. Int. J. Cogn. Ergon. **3**(3), 169–187 (1999)
20. Gevins, A., Smith, M.E.: Neurophysiological measures of cognitive workload during human-computer interaction. Theor. Issues Ergon. Sci. **4**(1–2), 113–131 (2003)
21. Haarmann, A., Boucsein, W., Schaefer, F.: Combining electrodermal responses and cardiovascular measures for probing adaptive automation during simulated flight. Appl. Ergon. **40**(6), 1026–1040 (2009)
22. Wilson, G.F., Russell, C.A.: Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. Hum. Factors: J. Hum. Factors Ergon. Soc. **49**(6), 1005–1018 (2007)
23. Lau, N., Jamieson, G.A., Skraaning Jr., G.: Situation awareness in process control: a fresh look. In: Proceedings of the 8th American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation & Control and Human-Machine Interface Technologies (NPIC & HMIT), San Diego, CA, USA (2012)
24. Rayner, K.: The 35th Sir Frederick Bartlett Lecture: eye movements and attention in reading, scene perception, and visual search. Q. J. Exp. Psychol. **62**(8), 1457–1506 (2009)
25. Scheiter, K., Van Gog, T.: Using eye tracking in applied research to study and stimulate the processing of information from multi-representational sources. Appl. Cogn. Psychol. **23**(9), 1209–1214 (2009)

26. Ahlstrom, U., Friedman-Berg, F.J.: Using eye movement activity as a correlate of cognitive workload. Int. J. Ind. Ergon. **36**(7), 623–636 (2006)
27. 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**, 58–75 (2014)
28. Hankins, T.C., Wilson, G.F.: A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. Aviat. Space Environ. Med. **69**(4), 360–367 (1998)
29. Kramer, A.K.: Physiological metrics of mental workload: a review of recent progress. In: Multiple-task Performance, pp. 279–328 (1991)
30. Marquart, G., Cabrall, C., de Winter, J.: Review of eye-related measures of drivers' mental workload. Proc. Manuf. **3**(Suppl. C), 2854–2861 (2015)
31. 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)
32. Wilson, G.F., Russell, C.A.: Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Hum. Factors **45**(4), 635–644 (2003)
33. Lin, Y., Zhang, W.J., Watson, L.G.: Using eye movement parameters for evaluating human–machine interface frameworks under normal control operation and fault detection situations. Int. J. Hum. Comput. Stud. **59**(6), 837–873 (2003)
34. Poole, A., Ball, L.J.: Eye tracking in HCI and usability research. Encycl. Hum. Comput. Interact. **1**, 211–219 (2006)
35. Just, M.A., Carpenter, P.A.: Eye fixations and cognitive processes. Cogn. Psychol. **8**(4), 441–480 (1976)
36. Matin, E.: Saccadic suppression: a review and an analysis. Psychol. Bull. **81**(12), 899 (1974)
37. Campbell, F.W., Wurtz, R.H.: Saccadic omission: why we do not see a grey-out during a saccadic eye movement. Vis. Res. **18**(10), 1297–1303 (1978)
38. Loftus, G.R.: Eye fixations and recognition memory for pictures. Cogn. Psychol. **3**(4), 525–551 (1972)
39. Kovesdi, C.R., Rice, B.C., Bower, G.R., Spielman, Z.A., Hill, R.A., LeBlanc, K.L.: Measuring human performance in simulated nuclear power plant control rooms using eye tracking. Idaho National Lab. (INL), Idaho Falls, ID (United States), INL/EXT–15-37311, November 2015
40. Palinko, O., Kun, A.L., Shyrokov, A., Heeman, P.: Estimating cognitive load using remote eye tracking in a driving simulator. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, pp. 141–144 (2010)
41. Yurko, Y.Y., Scerbo, M.W., Prabhu, A.S., Acker, C.E., Stefanidis, D.: Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool. Simul. Healthc. **5**(5), 267–271 (2010)
42. Cao, A., Chintamani, K.K., Pandya, A.K., Ellis, R.D.: NASA TLX: software for assessing subjective mental workload. Behav. Res. Methods **41**(1), 113–117 (2009)
43. Zhang, X., Mahadevan, S., Lau, N., Weinger, M.B.: Multi-source information fusion to assess control room operator performance. Reliab. Eng. Syst. Saf. (2018)

Roster and Air Traffic Controller's Situation Awareness

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

[1] Irish Aviation Authority, Dublin, Ireland
PETER.KEARNEY@IAA.ie
[2] Safety and Accident Investigation Centre, Cranfield University, Cranfield, UK

Abstract. Fatigue has been identified as a leading contributor to incidents and accidents within high-risk industries, in particular, the aviation sector. Traditional approaches used to mitigate fatigue, predominantly for air traffic controllers and flight crew, have largely focused on duty time limitations. FRMS is a data-driven means of continuously monitoring and managing fatigue-related safety risks, based on scientific principles and knowledge as well as operational experience, which aim to ensure relevant personnel are performing at adequate levels of alertness. It allows operators to adapt policies, procedures and practices to the specific conditions that create fatigue in a particular aviation setting. The ages of 36 participants in this study were from 23 years to 58 years old. The results demonstrated that ATCO's alertness indicate as functioning at a high level of alertness between 9 am and 5 pm during working hours. ATCOs' alertness levels deteriorated between working day-1 to day-5. Furthermore, there are significant differences over the 8 working hours per day among 5 working days. Particularly, the 6th, 7th and 8th working hours demonstrate much worsened alertness on day-4 and day-5. The pro-active approach of FRMS is to increase ATCOs fatigue resilience to cope with demanding situations while ATCOs are on position and while resting on breaks. FRMS is data driven, and data is collected from the operation, fatigue management decisions are made against this data, and the measurements that are required can be identified and implemented to improve the safety of operations.

Keywords: Air traffic management · Fatigue Risk Management · Human performance · Safety Management System · Situation awareness

1 Introduction

Fatigue Risk Management Systems (FRMS) came into being in the context of outcome-based regulation and Safety Management System (SMS) in the 1990s. Regulators and operators had concerns about the human and financial costs associated with fatigue and began to identify fatigue as another risk that could be managed using a systematic approach. Fatigue Risk Management progressively crystalized and today can be defined as explicit and comprehensive processes for measuring, mitigating and managing the actual fatigue risk to which an organization is exposed [1]. Factors that cause fatigue, such as circadian rhythm, sleep homeostasis and task-related influences have been demonstrated to have negative impacts to pilot's performance, thus creating

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 66–75, 2019.
https://doi.org/10.1007/978-3-030-22507-0_6

safety concerns in the aviation domain. In order to manage adverse consequences of fatigue, airlines have to implement fatigue risk management systems for flight crew and flight attendants. An FRMS, which is a scientifically-based on data-driven process, represents an alternative to the traditional prescriptive hours of work limitations. It manages employee fatigue in a flexible manner appropriate to the level of risk exposure and the nature of the operation. However, there are many FRMS variations for airlines pilots but not many for air traffic controllers. Therefore there is a need to develop an FRMS to optimize ATCOs' performance and increasing ATCOs' wellbeing.

1.1 Understanding the Impacts of Fatigue

Fatigue is a common complaint with a prevalence between 6.0 and 7.5% in Britain and the United States. A cross-sectional survey of United States workers found the two-week period prevalence of fatigue to be 38%, with an estimated annual cost to employers exceeding $136 billion in lost productive work time [2]. Fatigue has been identified as a leading contributor to incidents and accidents within high-risk industries, in particular, the aviation sector. The traditional approaches applied by regulators to mitigate fatigue, predominantly for air traffic controllers and flight crew, have largely focused on duty time limitations. However, these were often based more on the outcome of industrial negotiations, rather than being supported by scientific research. Fatigue is a complex physiological and psychological entity, it has negative affects to human operators in different ways and is influenced by numerous inputs, often outside the direct managerial responsibility or control of the employer. After some significant accidents, there has been a move to change the focus of fatigues analysis and mitigation towards managing the risk related to fatigue in a more scientific and evidence-based way.

Although fatigue is understood to have led to many aviation incidents and accidents, quantifying the exact share of fatigue as per these events is problematic because of the challenge of collecting hard evidence. Having this highlighted, fatigue is undoubtedly well established as a causal factor for many safety-related events. The negative effects of fatigue contribute to a general reduction in human performance, overall health drawbacks, and other social implications at the far end. Human operators experiencing fatigue may suffer from decreased decision-making skills, memory, judgment, reaction time and situational awareness. Compounding this is the fact that the effect of fatigue means that none of these symptoms may be apparent to the sufferer. Fatigue as a major symptom is found in all populations and is associated with multiple factors. Fatigue can be manifested as difficulty or inability initiating activity (perception of generalized weakness); reduced capacity maintaining activity (easy fatigability); and difficulty with concentration, memory, and emotional stability (mental fatigue) [3]. Duration of fatigue can be recent (less than one month), prolonged (more than one month), or chronic (over six months). The presence of chronic fatigue does not necessarily imply the presence of systemic exertion intolerance disease (SEID), also known as chronic fatigue syndrome (CFS).

1.2 Fatigue Risk Management in Flight Operations

FRMS is a data-driven means of continuously monitoring and managing fatigue-related safety risks, based on scientific principles and knowledge as well as operational experience, which aim to ensure relevant personnel are performing at adequate levels of alertness. It allows operators to adapt policies, procedures and practices to the specific conditions that create fatigue in a particular aviation setting. Operators may tailor their FRMS to unique operational demands and focus on fatigue mitigation strategies that are within their specific operational environment. As in Safety Management System, the FRMS relies on the concept of an "effective reporting culture and active involvement of all stakeholders where individual personnel have been trained and are constantly encouraged to report hazards whenever observed in the operational environment [4].

Fatigue, psychosocial workload and insufficient sleep have been recognized as a major concern of increased work intensity amongst working populations. Changes in the global economy and working life have increased the speed of business processes and the emergence of an increasingly '24/7 society' [5]. In addition, the need to increase work force flexibility and productivity has lengthened the average work day, shortened average recovery times and increased the irregularity of start and finish times. Indeed, fatigue is a common, almost universal feature of modern life. The effects of fatigue can vary but are best viewed as a continuum, ranging from mild, infrequent complaints to severe, disabling manifestations including burnout, overstrain, or chronic fatigue syndrome. The influence of fatigue on reduced individual performance that leads to incidents and accidents is well documented. NASA Ames Research Centre considers the role of fatigue in accidents as the contributory or causal role that fatigue may play in an accident is often underestimated or potentially ignored [6]. The U.S. National Transportation Safety Board has continually listed fatigue as one of its 'most wanted' safety improvements since 1996. According to Rosekind, a fatigue specialist and NTSB Board Member who proposed that it's not like you can't make decisions, it's just that you make bad decisions [7]. Extrapolating this analogy to the wider airline community, means that all levels of staff are vulnerable to the negative effects of fatigue and are generally worst placed to identify the problem.

1.3 Fatigue Risk Trajectory

Scheduling factors and non-scheduling factors are two useful categories to frame causal factors of fatigue in aviation. Scheduling factors are primarily connected with the rest periods and working intervals experienced by flight crew [11]. Operator fatigue in high risk industries is increased by extended time awake and reduced prior rest. In addition, changes in time-zones can present complex interactions between circadian lows and fatigue, further degrading performance [8]. In aviation, some of the present rules or proposed modifications of rules are in conflict with one or more of these factors [9]. The operational time limitations review panel organized by EASA [10] documented the key factors that can cause fatigue for flight crew members. There are multiple layers that precede a fatigue-related incident which are identifiable hazards and controls. An effective FRMS should attempt to manage each layer of risk as shown as Fig. 1. Dawson has described a multi-layered system of defenses based on assessing fatigue hazard and control mechanism [5]:

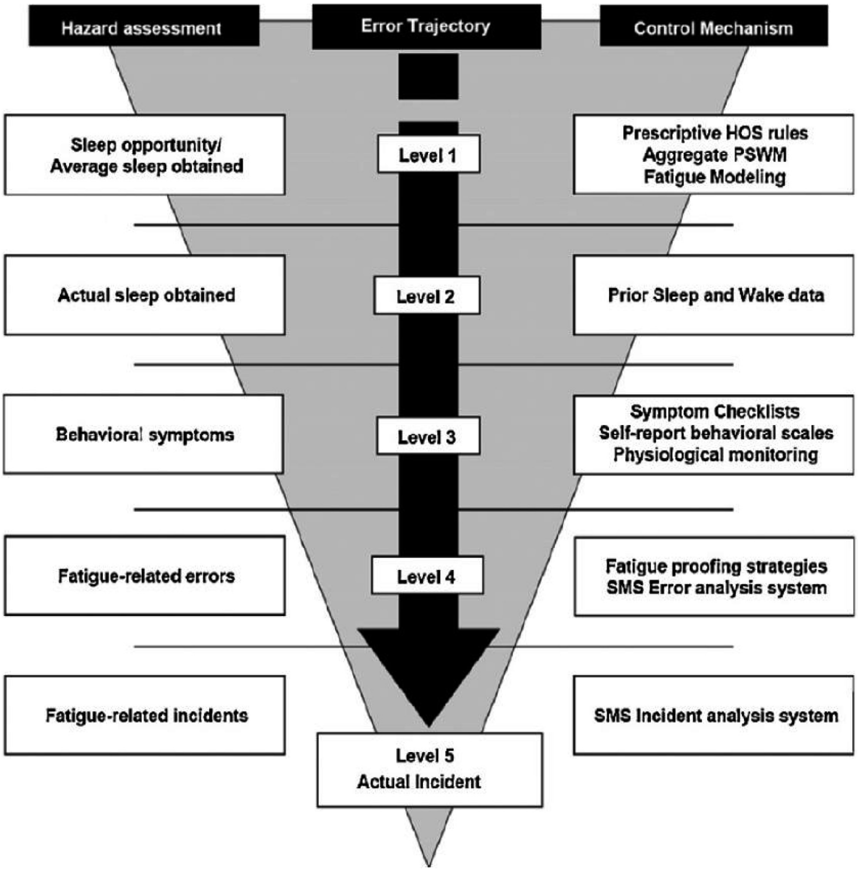


Fig. 1. Fatigue risk trajectory [13]

Level 1. Sleep opportunity afforded employees by the schedule.
Level 2. Actual prior sleep-wake behaviour experienced by the individual.
Level 3. Signs and symptoms of fatigue experienced by the individual.
Level 4. Nature, extent and preventability of fatigue-related errors.
Level 5. Fatigue-related incidents as organizational learning opportunities.

FRMS includes an accountable manager, who is ultimately accountable for fatigue risk, and it needs to exist within a just culture in which employees and management trust one another and information about fatigue is openly reported. A primary reason why FRMS is supposed to deliver enhanced protection against fatigue risk is because it measures actual risk and establishes tailored controls to mitigate or eliminate risks. Sleep is a key variable when considering the cause of fatigue in its most general sense [11]. The daily quantity of sleep required varies from one individual to the next, but on average it is eight hours [12].

2 Method

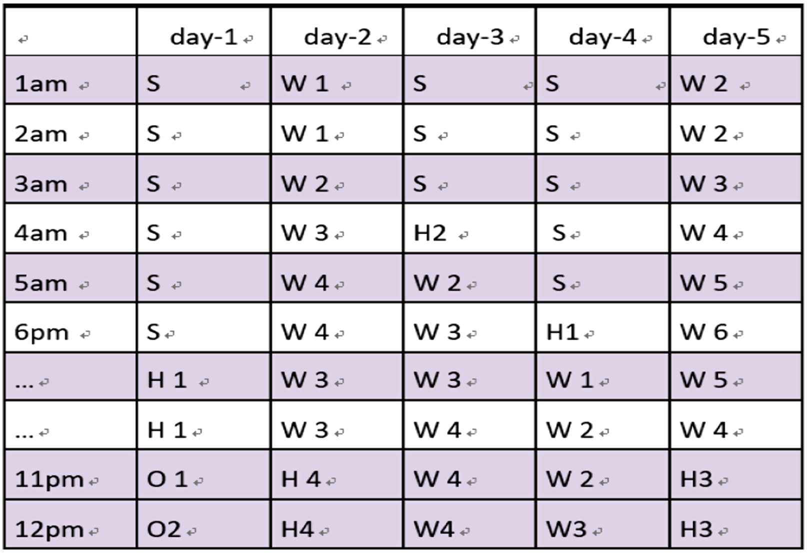
2.1 Participants

36 ATCOs participated in this research including 7 female and 29 male controllers. The ages of participants were from 23 years to 58 years old. The experiment process was reviewed and approved by the Research Ethics committee. The research objective is to offer the best rostering for ATCOs, and to provide the best approach for Fatigue Risk Management.

2.2 Alertness Rating Scale

The alertness rating scales start from 1 (the highest alertness) to 7 (the lowest alertness) contained 24 h a day for 8 days. The alertness can be divided into seven levels from 1 (the highest alert) to 7 (the lowest alertness) based on ATCO's experience while working on the position, meeting, taking breaks, at home including domestic activities such as watching TV, gardening, cleaning, cooking…; and other activities including social activities such as; gym, socializing, and exercising (Table 1). It is not required to indicate the alertness level for sleep (Table 1). This is a quick and easy way to assess ATCO's alertness levels while ATCOs are working and reflect the living patterns of day-to-day activities. This research only focused on ATCO's alertness level during their working hours on duty, as ATCO's alertness level reflect to situation awareness and therefore the potential impact on aviation safety.

Table 1. Example of alertness rating scales in ATM



	day-1	day-2	day-3	day-4	day-5
1am	S	W 1	S	S	W 2
2am	S	W 1	S	S	W 2
3am	S	W 2	S	S	W 3
4am	S	W 3	H2	S	W 4
5am	S	W 4	W 2	S	W 5
6pm	S	W 4	W 3	H1	W 6
…	H 1	W 3	W 3	W 1	W 5
…	H 1	W 3	W 4	W 2	W 4
11pm	O 1	H 4	W 4	W 2	H3
12pm	O2	H4	W4	W3	H3

2.3 Research Design

The participants were invited individually to a meeting room to receive instructions of how to use the alertness rating scales (Table 1) including 24 h a day across 5 shifts of working days in total. ATCOs have to record their alertness levels for each hour either working (W), social activities (S) or at home (H) using a7-points Likert scales. There is no requirement to indicate sleep (S) on the Likert scale, as the hours of sleep are considered as no alertness. Furthermore, ATCOs were encouraged to provide their feedback to develop the best practice of fatigue risk management to the project manager. There are demographical diversities of participants including male vs female, experienced ATCOs vs novice ATCOs, married vs not married, with children vs without children etc. It is a form of qualitative research consisting of interviews in which ATCOs been asked about their perceptions, opinions, beliefs, and attitudes towards the policy, procedures, training program and mitigation strategy regarding roster and fatigue risk management.

3 Result and Discussion

Participants' level of alertness is related to situation awareness on the controller working position. The collected data included 24 h among 5 shift working days shown as Fig. 2. The scheduled working hours over the 5 shift days are shown as following, day-1 from 13:30 to 21:30; day-2 from 14:30 to 22:30; day-3 from 07:45 to 16:00; day-4 from 06:30 to 13:30; and day-5 from 22:30 to 06:30. Scheduling factor is a useful category with which to frame causal factors of fatigue in aviation domain. Scheduling factors are primarily connected with the rest periods and working intervals experienced by operators [11]. Operator's fatigue in high risk industries is increased by extended time awake and reduced prior rest. In addition, changes in time-zones can present complex interactions between circadian and fatigue, further degrading situation awareness and human performance [8].

3.1 Roster Impacts to ATCO's Situation Awareness

The results demonstrated that ATCO's alertness is at a high level of alertness between 9 am and 5 pm during working hours (Fig. 2). A sample of these results further emphasized influences of time of day, time on duty, the complexity of tasks in one shift, the timing of sleep prior to duty starting, and effect of consecutive late finishes on fatigue. It was found that the effect of time of day was highly significant, and closely reproduced the trends observed under laboratory conditions, with lowest levels of alertness in the late night and early morning (23:00–05:00). The changes with time on duty were also highly significant, and this is critical factor on the design of roster for ATCOs.

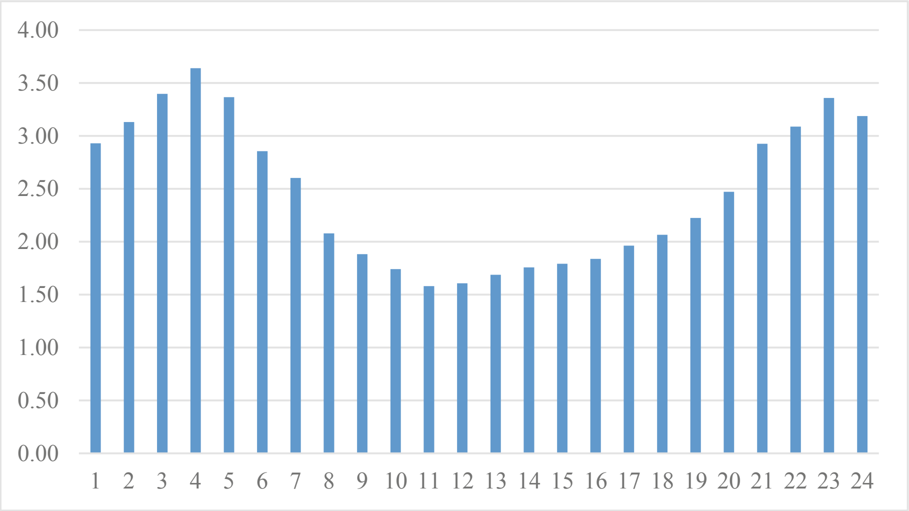
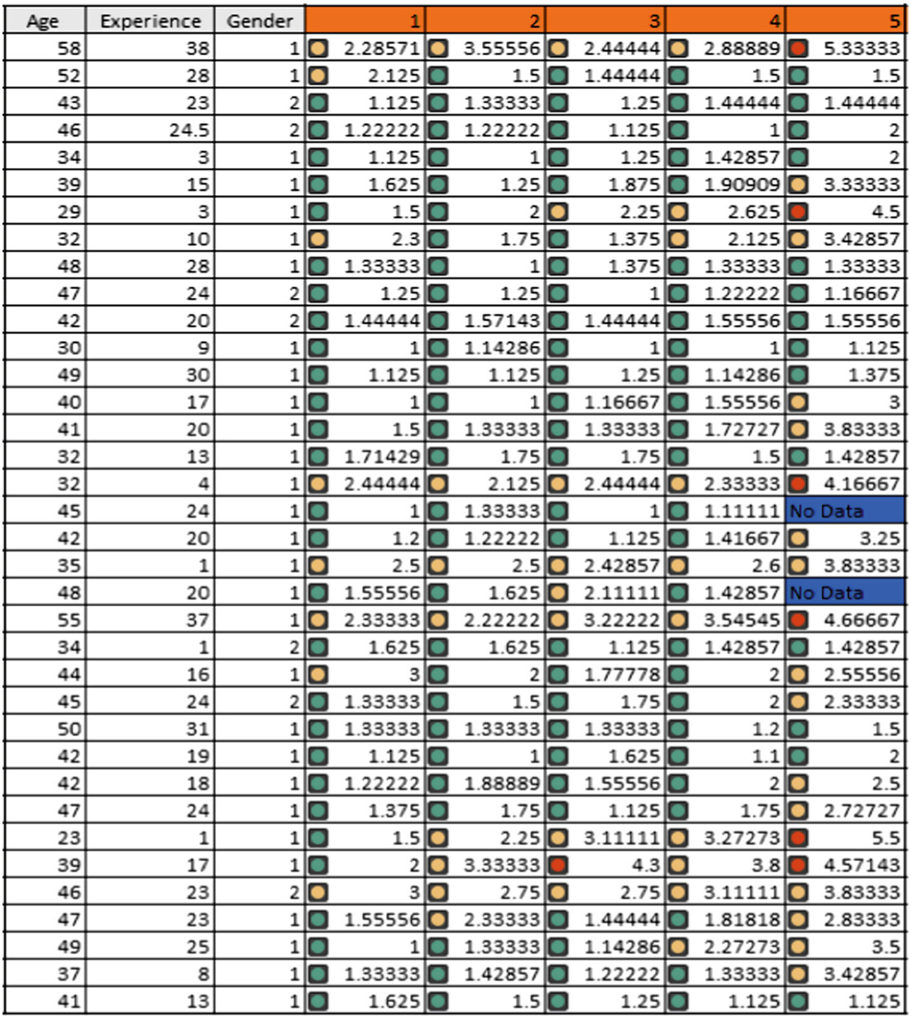


Fig. 2. The fluctuation of ATCO's alertness level among 24 h (1 indicates the highest alertness, 7 indicates the lowest alertness)

3.2 ATCO's Alertness Levels Among Shift Works

ATCOs' alertness levels deteriorated from working day-1 to day-5. Furthermore, there were significant differences during the 8 working hours among the 5 working days. Particularly, the 6th, 7th and 8th working hours, where alertness levels are much worse on day-4 and day-5 (Table 2). This is a safety concern as the rest time between day-4 and day-5 is only 9 h which does not provide sufficient sleep to maintain high alertness for high quality monitoring performance. Moreover, the working hours on both day-3 and day-5 include early morning and late night duty commencements. Sleep loss, and the resultant fatigue, both acute and cumulative, increases the experience of sleepiness, tension, confusion and decreased vigor. As little as two to three hours sleep loss on a single night can produce measurable increases in fatigue and resulting performance impairment on a variety of tasks both in the lab and in real-world settings. Further reductions in sleep or extensions in wakefulness can produce increasing levels of performance impairment similar in nature to moderate alcohol impairment. At more than 40 h of wakefulness, the resulting cognitive impairment can be both profound and debilitating [11, 13]. Hartzler [9] demonstrated that 24 h of continuous wakefulness was associated with significant deterioration on measures of reasoning and vigilance. The dangers associated with this level of impairment are then compounded by the fact that fatigued individuals are typically unaware of how severely their performance has deteriorated and thus may believe that they are safe to perform their duty when they are not [14, 15].

Table 2. ATCO's alertness levels among 5 shifts of working days



Age	Experience	Gender	1	2	3	4	5
58	38	1	🟡 2.28571	🟡 3.55556	🟡 2.44444	🟡 2.88889	🔴 5.33333
52	28	1	🟡 2.125	🟢 1.5	🟢 1.44444	🟢 1.5	🟢 1.5
43	23	2	🟢 1.125	🟢 1.33333	🟢 1.25	🟢 1.44444	🟢 1.44444
46	24.5	2	🟢 1.22222	🟢 1.22222	🟢 1.125	🟢 1	🟢 2
34	3	1	🟢 1.125	🟢 1	🟢 1.25	🟢 1.42857	🟢 2
39	15	1	🟢 1.625	🟢 1.25	🟢 1.875	🟢 1.90909	🟡 3.33333
29	3	1	🟢 1.5	🟢 2	🟡 2.25	🟡 2.625	🔴 4.5
32	10	1	🟡 2.3	🟢 1.75	🟢 1.375	🟡 2.125	🟡 3.42857
48	28	1	🟢 1.33333	🟢 1	🟢 1.375	🟢 1.33333	🟢 1.33333
47	24	2	🟢 1.25	🟢 1.25	🟢 1	🟢 1.22222	🟢 1.16667
42	20	2	🟢 1.44444	🟢 1.57143	🟢 1.44444	🟢 1.55556	🟢 1.55556
30	9	1	🟢 1	🟢 1.14286	🟢 1	🟢 1	🟢 1.125
49	30	1	🟢 1.125	🟢 1.125	🟢 1.25	🟢 1.14286	🟢 1.375
40	17	1	🟢 1	🟢 1	🟢 1.16667	🟢 1.55556	🟡 3
41	20	1	🟢 1.5	🟢 1.33333	🟢 1.33333	🟢 1.72727	🟡 3.83333
32	13	1	🟢 1.71429	🟢 1.75	🟢 1.75	🟢 1.5	🟢 1.42857
32	4	1	🟡 2.44444	🟡 2.125	🟡 2.44444	🟡 2.33333	🔴 4.16667
45	24	1	🟢 1	🟢 1.33333	🟢 1	🟢 1.11111	No Data
42	20	1	🟢 1.2	🟢 1.22222	🟢 1.125	🟢 1.41667	🟡 3.25
35	1	1	🟡 2.5	🟡 2.5	🟡 2.42857	🟡 2.6	🟡 3.83333
48	20	1	🟢 1.55556	🟢 1.625	🟡 2.11111	🟢 1.42857	No Data
55	37	1	🟡 2.33333	🟡 2.22222	🟡 3.22222	🟡 3.54545	🔴 4.66667
34	1	2	🟢 1.625	🟢 1.625	🟢 1.125	🟢 1.42857	🟢 1.42857
44	16	1	🟡 3	🟢 2	🟢 1.77778	🟢 2	🟡 2.55556
45	24	2	🟢 1.33333	🟢 1.5	🟢 1.75	🟢 2	🟡 2.33333
50	31	1	🟢 1.33333	🟢 1.33333	🟢 1.33333	🟢 1.2	🟢 1.5
42	19	1	🟢 1.125	🟢 1	🟢 1.625	🟢 1.1	🟢 2
42	18	1	🟢 1.22222	🟢 1.88889	🟢 1.55556	🟢 2	🟡 2.5
47	24	1	🟢 1.375	🟢 1.75	🟢 1.125	🟢 1.75	🟡 2.72727
23	1	1	🟢 1.5	🟡 2.25	🟡 3.11111	🟡 3.27273	🔴 5.5
39	17	1	🟢 2	🟡 3.33333	🔴 4.3	🟡 3.8	🔴 4.57143
46	23	2	🟡 3	🟡 2.75	🟡 2.75	🟡 3.11111	🟡 3.83333
47	23	1	🟢 1.55556	🟡 2.33333	🟢 1.44444	🟢 1.81818	🟡 2.83333
49	25	1	🟢 1	🟢 1.33333	🟢 1.14286	🟡 2.27273	🟡 3.5
37	8	1	🟢 1.33333	🟢 1.42857	🟢 1.22222	🟢 1.33333	🟡 3.42857
41	13	1	🟢 1.625	🟢 1.5	🟢 1.25	🟢 1.125	🟢 1.125

(green ball indicates high level of alertness, yellow ball indicates middle level of alertness, red ball indicates low level of alertness)

4 Conclusion

The effects of fatigue can vary but are best viewed as a continuum, ranging from mild, infrequent complaints, to severe, disabling manifestations including burnout, overstrain, or chronic fatigue syndrome. In spite of the complex nature of fatigue, the operational implications are strikingly consistent across diverse types of air traffic controllers. Based on the findings of this current research, it is impossible to develop a roster to satisfy all controllers, as each individual ATCO has differing preferences for a roster schedule catering to different genders, ages, marriage, family sizes and patterns of life. The operational demands in air traffic management continue to change in response to changes in volume of aircraft, new technology (remote tower) and commercial pressures (cost-efficiency), however human physiology remains unchanged. Both prescriptive fatigue management regulations and FRMS represent an opportunity to use advances in scientific understanding of human physiology to better address fatigue risks for ATCOs. The research has identified some concerns regarding the interval times between day-4 and day-5. Arguably, there is no evidence to suggest that the roster has directly affected the safety of service delivery over the duration of the established roster within the organization. The pro-active approach is to increase ATCOs fatigue resilience to cope with demanding situations while ATCOs are on the position and while resting on breaks. FRMS is data driven, and data are collected from the operation, fatigue management decisions are made against this data, and the measurements that are required can be identified and implemented to improve the safety of operations.

References

1. Holmes, A., Al-Bayat, S., Hilditch, C., Bourgeois-Bougrine, S.: Sleep and sleepiness during an ultralong-range flight operation between the Middle East and United States. Accid. Anal. Prev. **45**, 27–31 (2012). https://doi.org/10.1016/j.aap.2011.09.021
2. Ricci, J.A., Chee, E., Lorandeau, A.L., Berger, J.: Fatigue in the US workforce: prevalence and implications for lost productive work time. J. Occup. Environ. Med. **49**(1), 1–10 (2007). https://doi.org/10.1097/01.jom.0000249782.60321.2a
3. Yennurajalingam, S., Bruera, E.: Palliative management of fatigue at the close of life: "it feels like my body is just worn out". JAMA **297**(3), 295–304 (2007). https://doi.org/10.1001/jama.298.2.217
4. FAA: Flight crew member duty and rest requirements. US DOT FAA Docket No. FAA-2009-1093; Amendment Nos. 117-1, 119-16, 121-357, December 2011
5. Dawson, D., Noy, Y.I., Härmä, M., Åkerstedt, T., Belenky, G.: Modelling fatigue and the use of fatigue models in work settings. Accid. Anal. Prev. **43**(2), 549–564 (2011). https://doi.org/10.1016/j.aap.2009.12.030
6. Rosekind, M.R., Gregory, K.B., Co, E.L., Miller, D.L., Dinges, D.F.: Crew factors in flight operations XII: a survey of sleep quantity and quality in on-board crew rest facilities (2000)
7. Mark, R.P.: NTSB's Rosekind Warns of Pilot Fatigue and Sleep Problems at HeliExpo Safety Seminar (2012). http://www.ainonline.com/aviation-news/hai-convention-news/2012-02-12/ntsbs-rosekind-warns-pilot-fatigue-ar_isi-g_egp-problems-heliexpo-safety-,seminar. Accessed 29 Nov 2012
8. Gander, P., van den Berg, M., Mulrine, H., Signal, L., Mangie, J.: Circadian adaptation of airline pilots during extended duration operations between the USA and Asia. Chronobiol. Int. **30**(8), 963–972 (2013). https://doi.org/10.3109/07420528.2013.790042
9. Hartzler, B.M.: Fatigue on the flight deck: the consequences of sleep loss and the benefits of napping. Accid. Anal. Prev. **62**, 309–318 (2014). https://doi.org/10.1016/j.aap.2013.10.010
10. EASA: Scientific and medical evaluation of flight time limitations. EASA Final Report TS. EASA.2007.OP.8 (2008)

11. Dawson, D., McCulloch, K.: Managing fatigue: it's about sleep. Sleep Med. Rev. **9**(5), 365–380 (2005). https://doi.org/10.1016/j.smrv.2005.03.002
12. Van Dongen, H.P.: Shift work and inter-individual differences in sleep and sleepiness. Chronobiol. Int. **23**(6), 1139–1147 (2006). https://doi.org/10.1080/07420520601100971
13. Dawson, D., Chapman, J., Thomas, M.J.: Fatigue-proofing: a new approach to reducing fatigue-related risk using the principles of error management. Sleep Med. Rev. **16**(2), 167–175 (2012). https://doi.org/10.1016/j.smrv.2011.05.004
14. Banks, S.: Behavioral and physiological consequences of sleep restriction. J. Clin. Sleep Med. **3**(05), 519–528 (2007). https://doi.org/10.1055/s-0029-1237117.Neurocognitive
15. Durmer, J.S., Dinges, D.F.: Neurocognitive consequences of sleep deprivation. Paper Presented at the Seminars in Neurology (2005)

Socio-Technical Safety Investigations in Healthcare – Investigating Human Performance in Modern High Reliability Sector Organizations

Pete McCarthy[1,2](✉) and Andrew Blackie[1,2]

[1] Safety and Accident Investigation Centre, Cranfield University, Cranfield, UK
pete.mccarthy@cranfield.ac.uk
[2] ABRIS Consulting Ltd., Glasgow, UK

Abstract. The introduction of the Healthcare Safety Investigation Branch into the National Health Service (NHS) in England is a world first, independent, not for blame investigation approach for healthcare. These investigations are conducted in an environment which has vastly varying levels of socio-technical complexity across a wide geographical region of the United Kingdom (UK) and across Trusts, departments and specialist disciplines. At the heart of this system are the healthcare workers who constantly balance resource to ensure patient safety is maintained to the highest levels. Embedded in a socio-technical system, the human contribution is often providing the adaptability which makes the system work. Historically if patient safety was compromised, or an unexpected outcome occurred it was the human contribution which was scrutinized, often with a view to disciplinary or punitive action in order to prevent recurrence. A more modern approach to system thinking guides us to see the human contribution as only one element of a socio-technical system and possibly the richest source of evidence for fully understanding any event. This pilot study has identified the perceived qualities deemed most valuable for healthcare safety investigators for whom the investigation of human performance will be key to understanding the majority of patient safety events they respond to. Non-technical skills including communication, Emotional Intelligence, resilience and empathy were ranked above the clinical or technical skills as more important for the individual investigator conducting investigation in healthcare. This is dependent on the clinical and technical expertise being available at a team level to the individual investigator. The initial findings are interesting in that they appear to indicate that as the environment is becoming increasingly socio-technically complex, it is the softer, non-technical (human-centered) skills that are required to understand narrative and context when unexpected outcomes occur in the healthcare setting.

Keywords: Healthcare · Safety investigation · Socio-technical · Human performance

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 76–91, 2019.
https://doi.org/10.1007/978-3-030-22507-0_7

1 Introduction

The fast pace of technological advancement in society is clearly evident across many parts of the world in the 21st century. Over the past 80 to 100 years significant industries including commercial aviation, Air Traffic Control (ATC), healthcare and patient safety, power generation, financial markets and food production have become almost completely reliant on technology for their day to day transactions and management. As these industries have developed, the only real constant throughout has been the presence and influence of humans somewhere in the system. The human contribution is integral to making the overall system work and will include advancement and management of the technology, including the initial development of processes and systems to manage the Human Machine Interface (HMI), or Human Computer Interface (HCI). The working environments or work space that many of these humans occupy has morphed, from an almost simplistic, mechanistic workplace in to a complex socio-technical environment where humans are embedded in systems as agents alongside the technology (Stanton et al. 2010); the human while interacting with the technology is also often balancing resources, time, finances and even safety to provide flexibility and adaptability, this ensures the goals of the organisation are met (Hollnagel 2009). Such systems, composed of human agents and technical artefacts, are often embedded within complex social structures such as the organisational goals, policies and culture, economic, legal, political and environmental elements (Qureshi 2007). Socio-technical theory implies that human agents and social institutions are integral parts of the technical systems, and that the attainment of organisational objectives are not met by the optimisation of the technical system, but by the joint optimisation of the technical and social aspects (Trist and Bamforth 1951). Healthcare and patient safety, which is the main topic addressed in this paper is a good example of a modern complex socio-technical system. The technological artefacts (life support systems, ambulances, staffing and management technologies, robotic surgery, scanners, smart phones, tablets, communication systems and electronic health records for example) all play an essential role alongside the human agents in the functioning of the system as a whole.

The Socio-technical environment in healthcare is not standard or easily defined across all of the many specialist areas and disciplines - some of the specialist areas within healthcare might be considered to be very tightly coupled and very complex, whereas other might be loosely coupled and much simpler to describe. (Hollnagel 2009). The social/technical environment exists across all of healthcare however, therefore the behaviours and interactions/interventions of the human agents are key to the overall day to day management of this system. As technology became integral to many of the industries already mentioned (including healthcare), rules and procedures had to be developed to give order and structure to the environment and guide the tasks or processes, the aim being to improve productivity and achieve an outcome. More latterly in these industries we now see risk management processes employed in order to enhance safety, and safety regulation then appears alongside efficiency and thoroughness as competing goals (Hollnagel 2009). Operating procedures, rules and even laws were designed and implemented to direct the human and technical contribution, these procedures and rules at the local level were often introduced by management and decision makers far removed from the actual work being done, resulting over time in a clear gap between work as imagined by the management and designers, and work as done by the workforce closest to the day to day activity (Snook 2000). This gap also often being exasperated by rules, procedures and laws imposed from outside of the organisation i.e. government, regulators and professional bodies set up to represent and monitor professional standards within certain disciplines.

2 Background

When accidents occur in the modern socio-technical environment, they often occur as a result of the normal and expected interaction between the humans, the technology, the procedures, the environment and the equipment. An interaction which normally does not result in any negative outcome has, despite outwardly appearing to be the same as previous interactions, resulted in harm. Traditional accident modeling approaches are not adequate to analyse accidents that occur in modern socio- technical systems, where accident causation is not the result of an individual component failure or human error (Qureshi 2007).

In a time preceding the complex environment we now operate in, the cause of any mishap was often thought to be simple and clear; something mechanical or technical broke, a worker was negligent or was not following rules and procedures or was criminal or malicious. An investigation would quickly find the root cause. Equipment could be mended, replaced or subjected to better design or maintenance regimes. The worker could be dealt with through, retraining, blaming, shaming, discipline or dismissal. For the investigator in these early days the method of investigation was also quite simplistic. The investigator was reliant upon the accident causation models, theories or approaches available at the time, these models were devised from academic research for use in the applied setting, based upon the known complexity at that time. Accident models provide a conceptualisation of the characteristics of the accident, which typically show the relation between causes and effects. They explain why accidents occur and are used as techniques for risk assessment during system development and post hoc accident analysis to study the causes of the occurrence of an accident (Qureshi 2007).

Linear logic was employed to analyse the event (often in the early days only the immediate event) and once the component parts had been identified an almost mechanistic approach was employed in order to demonstrate cause and effect linkages. This cause and effect linkage was mostly temporal, i.e. Action A preceded Action B in time, which then resulted in the event under investigation. Action A may have been identified as the root cause and therefore recommendations would then be drafted to deal with whatever shortcomings were evident at Action A. These analysis methods, referred to as sequential methods evolved over time to include for example; “Root Cause Analysis (RCA)”, 5 Whys, Fault Tree Analysis (FTA), Event Tree Analysis (ETA), Sequentially Timed Event Plotting (STEP). In the time-line of safety investigations, this approach was dominant in many high reliability organisation investigations right up until the modern day, though most began a move away from these as their prescribed method by the early 1980’s. It should be noted however that one of the strengths (and weaknesses) of this type of linear approach, is the ability to create accurate timelines focused on the period proximal to the identified outcome, thus quickly establishing what happened and who or what was involved at that point. This approach does not adequately address how or why we arrived at the outcome - this being a crucial part of the event under analysis if we are to make recommendations to prevent recurrence. With this limitation in mind and following a number of serious industrial accidents with what appeared to be an organisational focus e.g. Three Mile Island, Bhopal and Chernobyl a new approach was required to adequately explain events. Simplistic linear methods did not necessarily capture the performance shaping factors of the workplace, organisation or environment and new approaches and thinking on behalf of the investigator would be required in order to do so. The need for more powerful ways of understanding accidents led to the class of epidemiological accident models, which began to gain in popularity in the 1980s.

This analysis approach came to prominence in the late 80s. One model aligned with this theory and conceptualised at this time is the well known but colloquially named Swiss Cheese Model, established by Professor James Reason (Reason 1990). Complex linear thinking with regard to accident investigation was considered to be the new approach required in order to get beyond the proximal event and begin to address the how and why of the accident. By working backwards and examining actions and events beyond the immediate, front-line we begin to address those elements which though not proximal to the outcome demonstrate a potential to affect the outcome. Reason referred to these as latent causal factors to differentiate them from the active areas previously focused on. Reason draws attention to "The significance of causal factors present in the system before an accident sequence actually begins… and all man-made systems contain potentially destructive agencies, like the pathogens within the human body".

Epidemiological models regard events leading to accidents as analogous to the spreading of a disease, i.e. as the outcome of a combination of factors, some manifest and some latent, that happen to exist together in space and time. Reason's (1990) Latent conditions including management practices or organisational culture are likened to resident pathogens and can lie dormant in a system for a long time. Reason referred to this approach as a total systems approach to safety although, as we will see with the systemic models that will follow in this timeline of approaches, the total system (outside of the organisation or institutions thought to be directly involved) may not have been adequately represented in these early "Epidemiological" types of investigations.

Reason based epidemiological approaches to investigation have been adapted by many industries and domains in order to best reflect the specific nuances of the organisations where an accident or serious incident has occurred - the Australian Transport Safety Bureau (ATSB) model (ATSB 2007) is one such approach others include HFACS (Wiegemann and Shappell 2003) and the Accident Route Matrix (Harris et al. 2016).

In a complex dynamic environment it is not possible to establish procedures for every possible condition, in particular for emergency, high risk, and unanticipated situations (Rasmussen 2007). Decision making and human activities are required to remain between the bounds of the workspace defined by administrative, functional and safety constraints. Rasmussen argues that in order to analyse a work domain's safety, it is important to identify the boundaries of safe operations and the dynamic forces that may cause the socio-technical system to migrate towards or cross these boundaries (Qureshi 2007). These boundaries; acceptable behaviour, safety regulation, economic failure and unacceptable workload form the edges of the space within which work as done (as opposed to work as imagined) is completed. Workers are constantly adapting their behaviour, processes and methods in order to meet the output requirement of their dynamic environment. Often this adaptation occurs within the safe space bounded by the four factors listed above, however if this adaptive practice crosses the boundaries of acceptable behaviour and safety regulation this may lead to a loss of control and an accident may be the result.

The sequential and epidemiological models have contributed to the understanding of accidents; however, they are not suitable to capture the complexities and dynamics of modern socio-technical systems. In contrast to these approaches, systemic models view accidents as emergent phenomena, which arise due to the complex and nonlinear interactions among system components (Qureshi 2007).

One common theme for all of the analysis concepts and approaches listed above, is that they set out to de-construct an event, situation, accident sequence or near miss, with the aim being to establish causal links. Some of these links will be proximal to the event while some, such as government policies and governing body direction, will be far removed. The investigator will plot the agents and artifacts involved (in their initial time-line of the incident), then aim to identify where any cause and effect, or lines of influence might be evident (findings based upon analysis). The most up to date systemic methods purport to be non-linear and complex in their approach, however there is still almost always a temporal order required to establish a cause and effect relationship. Is this cause and effect linkage even required in a socio-technical system in order to adequately de-construct the event?

If the aim of a modern safety investigation is not to apportion blame or liability but to prevent recurrence, learn lessons and make the system safer, is there any requirement to establish cause and effect linkages at all? A more pragmatic approach to understanding patient safety events in healthcare for example might be one of getting the whole story, understanding the complete context behind actions, decisions and behaviours in order to determine why people's actions made sense to them at the time, rather than isolating them in order to place them in some perceived order for understanding. We are then looking at human performance investigations whereby the mechanistic, simplistic and even the more complex epidemiological models may only serve as a start point for the investigators quest to provide robust recommendations in order to prevent recurrence, make the system safer and learn lessons across the whole of their industry and beyond.

This paper explores healthcare as a socio-technical environment in which the approach for analysing serious incidents and accidents is constantly evolving (in the UK this is currently on a month by month basis). The Healthcare Safety Investigation Branch in England is working to understand the socio-technical complexity of healthcare and they are taking a forward-thinking approach to the accident analysis of serious incidents and accidents in that domain. They are leading the world in the application of not for blame, independent safety investigation in this field. It is this approach and work by investigators in the UK which forms the evidence for this pilot study "Investigating Human Performance in Modern High Reliability Sector Organisations". A combination of many approaches is employed in healthcare in order to understand what, how and why serious incidents have occurred. Simple linear methods are used to determine what happened, more complex linear and systemic thinking helps to establish how and why, and the application of specific Human Performance Investigation methods, alongside clinical/technical approaches, are actively followed. Second and third victim considerations centre around staff, family members or other witnesses who may suffer or perceive to suffer trauma from the event under consideration. Early anecdotal evidence from investigations already concluded in the past 24 months in healthcare appear to support the move away from person centered investigations and a concentration on the proximal event and proximal actors. It is worthy of note that the systemic models at present across safety investigation tend to be restricted to theory and concept with regard to application, whereas the epidemiological models and approaches are actively being applied across industry including healthcare (HFACS, ARM, Maternity Investigation Matrix (MIM).

This research aims to take the concepts and theories around epidemiological and systemic models and create an applied approach, which follows traditional thinking only in the initial deconstruction of the event, but then goes further to consider:

- Context
- Narrative
- Positive action (not only negative causal path, but positive performance influence also)

This pilot study forms the basis of a much larger project which will continue over the next 24 months, and it will aim to provide a framework of key competencies, experience and knowledge for the healthcare safety investigator.

3 Research Question

- Is the Socio-Technical complexity of the healthcare environment adequately deconstructed for investigative purposes by the current accident analysis models in use across other High Reliability Organisations (HRO).
- Which skill sets (technical or non-technical) are most valuable to the investigator, conducting Human Performance investigations in healthcare.
- Are well-developed social skills or a well-developed social approach to investigation required for understanding patient safety occurrences and events in healthcare.

4 Method

4.1 Participants

This pilot study serves as the precursor to a much larger project planned for mapping the healthcare investigator competencies. The pilot has taken advantage of the recent requirement in the UK to qualify a large number of healthcare investigators (100+) over a relatively short period of time (18 months), and ensure these investigators are taking the same approach to "not for blame" independent and transparent investigation as employed by other State level investigators in the United Kingdom namely the Air, Rail and Marine Accidents Investigation Branches known as the AAIB, RAIB and MAIB respectively.

The aim of the pilot study was to capture what newly recruited investigators in the healthcare domain deem to be the most important personal qualities for them to demonstrate whilst conducting investigations. The perception being that their role will be centered around conducting a human performance investigation in the healthcare environment.

Investigators in this role understand their place within the wider Healthcare safety Investigation Branch, whereby clinical, technical Subject Matter Advisors (SMAs) are on hand to provide guidance with regard to specific clinical or technical issues. They clearly understood that they would however need to identify and understand technical and clinical factors present in their investigation and be able to recognise where these elements may sit within the context and narrative of the investigation.

The aim will be to revisit this cohort as they gain experience over the coming years in order to see how the theoretical competencies match the applied competencies over time.

More than 100 personnel have been selected for employment by the HSIB with varying levels of medical/clinical expertise, including some with no medical or clinical expertise at all. A key feature of their selection however has been with regard to their demonstrated (at interview) behaviours and attitudes which the management team perceive to be a good fit with the organization's currently perceived requirement of the healthcare investigation environment.

The participants (50) for the pilot study is almost 50% of the current population of specialist healthcare safety investigators. This is a high sample size particularly regarding the specialist nature of the investigator role being studied. All participants have undergone a rigorous interview process, they have completed a week of induction into the new Healthcare Safety Investigation Branch and are part way through their safety investigation training when the researcher has collected the data for this pilot study. Participants have at this stage a clear understanding of the aims and approaches of the organisation they have joined, the criteria by which an event requiring their attention as investigators will be triggered, the purpose of a safety investigation and some of the methods, tools and techniques available to the safety investigator.

Design and Procedure. One design has been used for the pilot study (initial generation of qualities of an investigator), though it is envisaged that two designs (generation of competencies and generation of a Hierarchal Task or Cognitive Task Analysis) will be required for the future (main studies).

Participants, having been selected for the role (as described above) were asked to first generate then rank the qualities of an investigator that they believe were most important for the type of investigation they understood they would be tasked with as soon as they had completed their training. Participants were given a short generic brief which introduced the concept of the qualities of a safety investigator. The socio-technical complexities across healthcare were discussed – their own generated understanding of this produced specific areas within health ranging from easier to describe, non-complex areas such as General Practice (GP) through to more complex difficult to describe departments like Emergency Departments (ED). A short exercise took place whereby they were asked to map these complexities across healthcare and then across their specific domain – maternity. This exercise was conducted in order to ensure each investigator was considering the full range of their investigator task domain, before addressing the research task. Participants were then split into groups of 4–5 individuals, and were allocated 20 min to discuss, generate, rank and agree upon up to 8 qualities they determine to be essential for their new role in healthcare. The participants were completely free at this stage to choose the descriptors they agreed reflected the best qualities required. They were not given a pre-determined list to rank as it is the researchers plan to use their list for future detailed studies involving more experienced (potentially the same investigators after they have concluded 10 or more investigations). Once complete, the investigators were asked to produce their ranking and discuss briefly the rationale for their list. All descriptors were collated and analysed across the groups for the prevalence of perceived importance.

5 Results

Figures 1 and 2 below shows an example of the complexity mapping exercise conducted by the participant groups. The groups were free to alter the language used to describe the complexity and they chose to move away from the Hollnagel descriptors of manageability (which they replaced with predictability) and tractability or coupling (which they replaced with interdependence).

The participants then generated their own list of descriptors for investigator qualities during this task. These were deemed essential to investigate human performance in the socio-technical areas identified above. These descriptors are:

- Communicator/Listener
- Team Player
- Empathy
- Integrity
- Resilient
- Approachable
- Compassion
- Credible
- Emotional Intelligence (EI)
- Curious
- Non-Judgemental
- Trustworthy
- Unbiased
- Self-aware
- Observant
- Kind

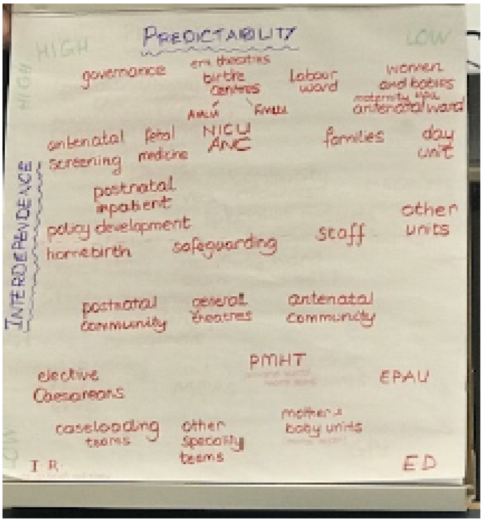


Fig. 1. Perceived socio-technical complexity across maternity

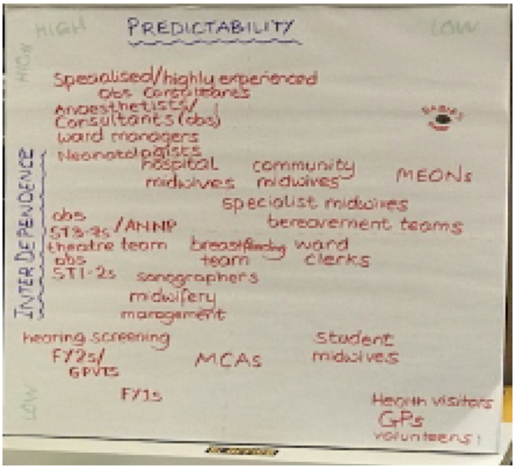


Fig. 2. Perceived socio-technical complexity across maternity

Tables 1 and 2 below show how the groups generated and then ranked these descriptors as small groups.

Table 1. Qualities ranked by cohort 1

Group 1	Group 2	Group 3	Group 4	Group 5
Good listener/communicator	Communicator	Integrity and honesty	Emotional intelligence	Good communicator
Kind/compassionate	Self-aware	Resilience	Empathy	Non-judgemental
Non-judgemental/objective	Trustworthy credible	Communicator	Communicator	Compassion ate
Approachable	Kind/emotional intelligence	Non-judgement al	Non judgemental	Kind
Dispassionate/self aware	Open and non judgemental	Curiosity	Approachable	Resilient
Inquisitive	Objective observer	Team-working	Kind	Independent
Resilient	Enquiring and analytical	Self awareness	Compassion ate	Open and honest
Independent	Resilient	Empathy	Trustworthy	Trustworthy
Open and honest				

Table 2. Qualities ranked by cohort 2

Group 1	Group 2	Group 3	Group 4	Group 5
Team player	Empathy	Team player	Emotional intelligence assertiveness	Communicator
Open minded	Patience	Impartial	Knowledge and experience	Open minded
Integrity	Communicator	Objective	Integrity	Compassionate
Non bias	Curiosity	Obsessive	Communicator	Emotional intelligence
Curiosity	Objectivity	Open minded	Flexibility	Kind
Humility	Integrity	Unbiased	Observant	Trustworthy
Thoroughness	Open minded	Patience	Leadership	Curious
Sense of judgement	Knowledge and skills	Good listener	Team player	Resilient
Compassion	Resourcefulness	Structured approach	Approachable	Credible

Once each cohort had completed the task, the combined ranking was analysed to check for prevalence of qualities identified. Figure 3 below captures this ranking.

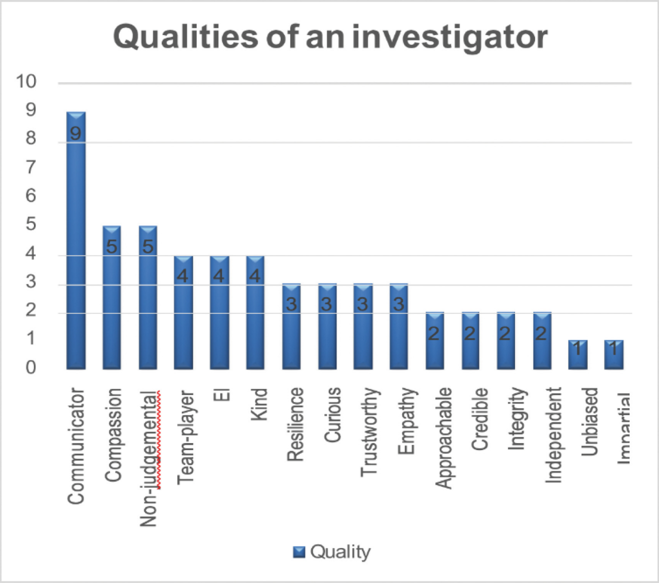


Fig. 3. Maps the qualities ranked by the participants

6 Analysis and Discussion

When considering the results from this pilot study, it is of particular interest to note the lack of identified qualities which might be categorised as either technical skills or those qualities perceived to be of a clinical nature. These initial findings are similar in many respects to the results found in the recent study of investigator competencies, by Nixon and Braithwaite (2018) "What do aircraft accident investigators do and what makes them good at it? Developing a competency framework for investigators using grounded theory". However it is of note that these healthcare professionals did not consider or report on organisational logistics, leadership or the practicality aspects of investigation when deciding upon the descriptors they deemed the most important qualities. It is, of course, possible that once they have had more exposure to investigation these rankings may change and further studies with this cohort will allow studies of intra-rater reliability over extended temporal periods to be undertaken.

Each cohort clearly demonstrated their comprehension of the socio-technical complexity across the domain they would be working in and there was clear understanding and comprehension of the impact the technical environment had on the frontline worker in healthcare on a daily basis. The researcher discussed in length after each exercise how the complexity of the environment impacted "work as done" by those performing tasks in healthcare. It was made clear that the delegates, whilst conducting the mapping exercise were doing so showing due consideration to the human in the loop, the technology behind their activity and their place as agents alongside these technical artefacts.

When discussing current approaches the methods and concepts of safety investigation analysis, spanning simple linear, complex linear and complex non-linear systemic methods, each cohort was satisfied that they understood where they might apply the different approaches to different areas of healthcare. Where low interdependence but high manageability or predictability was identified, they thought this to be quite simple and easy to describe, therefore more simple linear methods (RCA, 5 Whys) could be used. As Interdependence became higher and predictability remained high, then epidemiological approaches were well suited (HFACS, ARM, MIM). Where interdependence was high, but predictability was low, they saw this to be the most complex and difficult to describe areas which would require a much more systemic approach in order to analyse the complex socio-technical environment (ACCIMAP, STAMP). The delegates then quantified the quality descriptors as being vital to take these current methods even further in order to properly uncover human performance narrative and context.

It was apparent from discussion between the researcher and the delegates following each stage of the study detailed above that the technical/clinical environment in which investigations would take place was well understood. The delegates were content that they either possessed the expertise required to understand this element of the investigation, or they could call upon that expertise from within their local or wider team if required. It is of note that the qualities associated with this technical/clinical expertise did not feature in the qualities deemed to be most important in their new role as safety investigators. Though these skills are taught and trained they were not ranked at all, instead the skills sometimes referred to as soft skills or non-technical, non-taught such as, for example; Emotional Intelligence and Empathy were deemed more important.

From a human performance investigation perspective the results of this early pilot study opens up the prospect of further detailed research as to the perceived importance of these non-technical skills. In a not for blame safety investigation where the focus is on why and how an event occurred “what or who” are only important in order to complete the narrative to understand the proximal event. An ability to de-construct the event in order to understand the component parts may still be essential, and there are already adequate tools available for the investigator to do this. However, this deconstruction is no longer important as a means by which only the negative or problem areas on the direct causal pathway are mapped, instead the positive interactions need to be identified and captured also.

In order to capture the positive and negative interactions and map their significance, the investigator needs to engage with front line workers, family members, management, regulators, manufacturers and policy makers across healthcare. They need these agents to be open and honest with them in the understanding that the investigation is not for blame and that they seek only to make the system safer, prevent recurrence of harmful events and learn lessons. It might be said that this is the same across other high reliability domains (Aviation, ATC, Power generation), however the healthcare domain from this early research with investigators does appear to have a broader range of socio-technical complexity for the investigator to work with – with the human in the loop balancing the resources which appear far less constant and predictable than that experienced in other high reliability domains. The complexity varies from medical Trust to medical Trust, department to department, ward to ward, and the investigator needs to employ a range of skills which will enable them to map this environment and put any event into the context in which it occurred.

The skills deemed important by the delegates in this study (detailed below) are all non-technical (some of which it might be argued cannot be taught) yet are crucial to deconstructing an event or outcome in healthcare in order to fully understand the human contribution, add context and build a narrative of the event.

- Communicator/Listener
- Team Player
- Empathy
- Integrity
- Resilient
- Approachable
- Compassion
- Credible
- Emotional Intelligence (EI)
- Curious
- Non-Judgemental
- Trustworthy
- Unbiased
- Self-aware
- Observant
- Kind

Taking just the top 4 identified qualities:

Communication/Listening. From discussion following the exercise - This skill or quality was deemed to be key to the success of the healthcare safety investigation. Investigative interviewing was identified as a key method of gathering information and data for analysis and the delegates perceived that this is an area where their own personal abilities and qualities around communication and listening would be crucial to their task.

Compassion. From discussion following the exercise - The delegates were very conscious that staff and family members involved in any event were also second or third victims affected by the event under investigation, there was a general opinion among the delegates that this fact has been often overlooked in healthcare investigation historically. The investigation should learn from these second and third victims and put measures in place to protect or prevent future or further harm or sign-post for help if required. This also applied to other team member involved it the investigation.

Non-judgemental. From discussion following the exercise - The delegates believed from experience that investigations in their domain were historically seen as punitive and disciplinary in nature (though often the stated aim of the investigation was that of safety) and that, this was not in keeping with the type of investigation they sought to conduct. They determined that by taking a non-judgmental, not for blame approach to the investigation, they would eventually win the trust of the organisation – resulting in a more just culture across healthcare. Delegates were clear in their determination that it is the investigation that is not-for-blame but it is crucial that the culture remains just whereby accountability is understood, but a willingness to speak up when genuine "human error" or human (socio-technical) interaction conflicts are identified is maintained. This becomes the start point in identifying systemic issues where previously it would be the end point of a person-centered investigation.

Team-Player. From discussion following the exercise - The delegates were clear that they would need to rely on the skills across the team in order to best conduct their investigation, complete their analysis and provide credible, workable reports and recommendations. Not all of the delegates were from a clinical, technical background and though these skills were highly valued by the cohort as a whole, it was the dispersion of these skills across the teams that was determined to be of most value. Healthcare and the environment where those working in healthcare reside within the UK system has become increasingly socio-technically complex across all domains, departments and specialist areas. Some areas and disciplines are far more complex and some far less complex than others, but they all rely on the human in the loop (system) to balance resources, safety, performance and output. Millions of interactions, interventions and procedures are conducted each day across the whole system, and the adaptability of the human agent in this system is often considered to provide the underlying flexibility, adaptability, Quality Assurance (QA) and oversight to manage many conflicting priorities and deal with challenges and conflicts as they arise. On rare occasion this "work as done" or normal functioning of the dynamic system may result in a harmful outcome, on other occasions there may be moments of brilliance that save the day – but on the whole, the day to day reality is some positive and some negative socio-technical interactions provide for the normal day to day functioning of the system whereby the output standard meets and sometimes even exceeds that expected.

Of note at this pilot stage of the study is the perception of the delegates that as the environment becomes more socio-technically complex, it is the human "softer skills" that are required to fully understand the human interaction within this complexity. Technical know-how and a clear understanding of policies, processes, procedures and systems needs to be available to the team of investigators, but the personal and cultural context and narrative needs to be equally accessible. These latter elements can only be uncovered in the investigation by dealing at a personal level with those human agents embedded within the system. Deconstructing events using traditional methods (linear, epidemiological and systemic) are all helpful tools and approaches for the investigator, which should be maintained, as they will help represent the scenario and will give structure to the analysis. This pilot study demonstrates however that in healthcare it is not necessarily sufficient to only isolate cause and effect linkages, causal pathways or only the negative consequences and outcomes – instead all of the positive human, technical, system and environmental elements need to be captured also. This needs to be done at a local level on a case by case basis, it is not adequately captured by reporting systems. It is skillful human-centered, (taking the human perspective and narrative of the human), not person centered (whereby the human at the sharp end is deemed responsible for the outcome) human performance investigations which are required to completely understand the human contribution in the healthcare system.

The qualities identified by the delegates are all deemed necessary to allow the investigator to engage at a personal level with those involved in events deemed to meet the criteria for conducting a professional safety investigation. The delegates were clear that these personal qualities were most important but could not be isolated completely from either personal technical, clinical or process knowledge and experience – or at the very least the availability of these technical, clinical skill sets at a local team level. It is interesting at this pilot stage to consider whether all professional investigators working in healthcare need to demonstrate these personal “soft or non-technical” skills identified by these cohorts, or is it sufficient to have adequate numbers of team members able to so dispersed across the teams. Discussion following the research alluded to the concerns that this personal interaction with first, second and even third victims on a regular basis, might bring with it some emotional risks for the delegates and although emotional intelligence and resilience were clearly identified as key qualities for the investigator, it may be unkind to expose only a few team members to this potential trauma or risk. This interaction is required though to understand the full context and build the narrative, to allow for credible, measurable safety recommendations to be generated.

7 Conclusion and Recommendation

The purpose of this pilot study was to:

- Begin to map the complexity of the socio-technical environment in healthcare.
- List the qualities of an investigator deemed essential for working in this environment.
- Lay the groundwork for future studies once the current investigation branch (in its infancy at present) becomes established and more mature.

The initial findings are extremely interesting in that they appear to indicate that as the environment is becoming increasingly socio-technically complex, it is the softer, non-technical (human-centered) skills that are required to understand narrative and context when unexpected outcomes occur. This may be particular to healthcare, due to the perceived caring function of the system as a whole, or it may be an indication that in order to determine how and why workers take particular courses of action on a minute by minute, or second by second basis we have to build rapport and trust rather than display objective critical thinking. This objective critical thinking will be required at a team level when applying investigative judgement and expertise, but it will only come after the data and evidence has been gathered, which in a human performance investigation setting means interacting with people in a manner which most accurately reflects the potential trauma felt by those people. The next stage for this study will be to re-visit these cohorts once they have significant experience conducting investigations. At this time a comparison will be made against the pilot study results to determine how robust these initial findings are and to begin to map the competency framework for future investigators.

References

Australian Transport Safety Bureau. Analysis, Causality and Proof in Safety Investigations, ATSB Transport Safety Report Aviation Research and Analysis Report AR-2007-053

Harris, S.: Errors and accidents. In: Ernsting's Aviation and Space Medicine, pp. 707–722. Taylor and Francis Group, Boca Raton (2016)

Hollnagel, E.: The ETTO Principle: Efficiency-Thoroughness Trade-Off: Why Things that Go Right Sometimes Go Wrong. Ashgate, Aldershot (2009). ISBN 0-7546-7678-1

Nixon, J., Braithwaite, G.R.: What do aircraft accident investigators do and what makes them good at it? Developing a competency framework for investigators using grounded theory. Saf. Sci. **103**, 153–161 (2018)

Qureshi, Z.H.: A review of accident modelling approaches for complex socio-technical systems. In: Cant, T. (ed.) Proceedings of the Twelfth Australian Workshop on Safety Critical Systems and Software and Safety-Related Programmable Systems, (SCS 2007), vol. 86, pp. 47–59. Australian Computer Society, Inc., Darlinghurst (2007)

Rasmussen, J.: Risk management in a dynamic society: a modelling problem. Saf. Sci. **27**(2), 183–213 (2007). https://doi.org/10.1016/S0925-7535(97)00052-0

Reason, J.: Human Error. Cambridge University Press, New York (1990)

Snook, S.A.: Friendly Fire. Princeton University Press, Princeton (2000)

Stanton, N.A., Salmon, P.M., Walker, G.H., Jenkins, D.P.: Is situation awareness all in the mind? Theor. Issues Ergon. Sci. **11**(1–2), 29–40 (2010). https://doi.org/10.1080/14639220903009938

Trist, E.L., Bamforth, K.W.: Some social and psychological consequences of the Longwall method of coal-getting: an examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. Hum. Relat. **4**(1), 3–38 (1951). https://doi.org/10.1177/001872675100400101

Wiegemann, D.A., Shappell, S.A.: A Human Error Approach to Aviation Accident Analysis. Ashgate Publishing, Farnham (2003)

Research on Workload-Based Prediction and Evaluation Model in Power System

Caifang Peng, Zhen Wang(✉), Yanyu Lu, and Shan Fu

Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China
{pengcaifang, b2wz}@sjtu.edu.cn

Abstract. In the process of power grid dispatching, inappropriate workload may reduce dispatcher's work efficiency, and even lead to accidents. The purpose of this paper is to predict the current workload level of dispatchers, evaluate the current human risk level, and design a safe and reasonable work plan. In this paper, a multi-resource occupancy model based on VACVP (Visual Auditory Cognitive Voice Psychomotor) is proposed to predict workload. Meanwhile, a comprehensive evaluation model is proposed to validate the prediction model, which mainly uses PCA method to analyze the characteristics of physiological indicators such as heart rate, voice, movement to work out the actual workload. Finally, by means of time stamp alignment method, the workload results obtained from the two models are aligned and compared. Experimental results show that the workload predicted values obtained from VACVP workload prediction model are in line with the actual workload process. Furthermore, in a certain period of time, the trend of workload forecasting value is consistent with that of actual workload value, and the average workload data error is 0.6 grade, these ensure the validity and accuracy of the workload prediction model.

Keywords: VACVP multi-resource occupancy model · Physiological measurement · PCA · Comprehensive evaluation mode · Human workload

1 Introduction

At present, researchers have reached a general consensus that 60% to 90% of all system accidents, regardless of their field differences, can be attributed to human error. Especially in the process of power grid regulation and control, because it is not directly related to the operation of power production equipment, but the cognitive decision-making of power grid state by regulators, human factors have a particularly significant impact on the safe operation of power system.

Yerkes-Dodson rule [1] holds that there is an inverted U-shaped relationship between workload and performance. Moderate workload level can make performance reach its peak state, but too small or too large workload will reduce work efficiency, as Fig. 1 shows.

Appropriate workload intensity can improve the efficiency of operators, reduce the error rate of human risk, and significantly to the safe operation of the power grid.

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 92–105, 2019.
https://doi.org/10.1007/978-3-030-22507-0_8

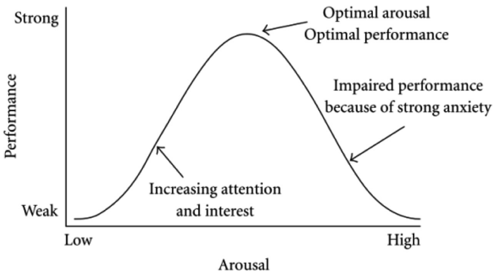


Fig. 1. Graph of Yerkes-Dodson law

Typical workload assessment methods include subjective evaluation, performance measurement and physiological measurement [2–4]. The subjective evaluation method, such as NASA-TLX, which cannot reflect the detailed situation of dispatchers at all times. Task performance measurement is based on the completion of the task, while in the grid system, he dispatching performance has delayed, because of the power dispatchers are not the actual operators of power facilities. Physiological measurement relies on the evaluation of physiological signals to assess workload. The results of this method are objective and reasonable to some extent, but it requires extended time and money.

This paper combines cognitive theory with system engineering method, uses Wicken’s and Yeh [5] multi-resource channel theory as shown in Fig. 2, which decomposes and divides the operator’s working process according to his behavior, and interprets the resource allocation relationship.

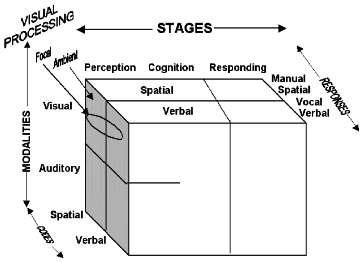


Fig. 2. Multiple resource model theory

On the basis of Wicken, McCracken and Adrich [6] proposed the VACP (Visual Auditory Cognitive Psychomotor) model, which takes the occupancy of multi-channel resources as the main index to evaluate the workload, but this method ignores the accumulation of time factors in the process of task execution.

In this paper, we proposal a VACVP (Visual Auditory Cognitive Voice Psychomotor) model, which includes five resource channels: visual channel, auditory channel, cognitive channel, voice channel and psychomotor channel. This model covers physiological and psychological workload evaluation, and can describe the workload intensity of power dispatchers comprehensively. What's more, it considers the time factor. According to Wickens [7], current workload can be judged by calculating the resource occupancy of each channel in a certain period of time. Therefore, we establish a comprehensive, continuous system suitable for serial and parallel tasks workload prediction model which based on multiple resource occupancy.

Besides, we proposal a comprehensive workload evaluation model based on physiological factor measurement to verifies the accuracy and availability of the VACVP workload prediction model. In this model, operators' physiological factors such as voice, behavior, psychology and action are objectively evaluated by camera, micro-phones and heart rate instrument. The actual workload value is mainly calculated by PCA, including feature extraction and weight calculation.

2 Methodology

2.1 Task

Multiple channel physiological data were collected from DTS anti-accident exercise scenario and daily work scenario of dispatching hall, respectively. In the dispatching hall, the daily work of dispatchers mainly includes record and monitor devices, which lasts for a long time (6–8 h) and has a large working area. In the DTS anti-accident scenario, its main characteristics are short duration, high workspace intensity, mainly dealing with accidents, and heavy mental workload. In this data acquisition method, it can reflect the work status of power dispatchers comprehensively.

2.2 Participants

Four power grid dispatchers were selected as the experimental subjects, with an average age of 33 years old and working life at about 6 years. The experimental subjects had no physical and psychological problems. They did not take tea, coffee and other psychoactive drugs before and during the experiment, in order to ensure the natural and good state in the experimental process.

2.3 Apparatus

In this study, the comprehensive evaluation model is based on physiological measurement, which need some measuring equipment as Fig. 3 shows. The camera is used to record the video in order to get the motion indicators, which has 140° wide angle of view, 1080P resolution, 60 Hz sampling frequency and high sensitivity. The speech and visual indicators are collected by the camera built in microphone. The heart rate indicator Mio Alpha heart rate watch obtains dispatcher's heart rate information.

Fig. 3. Camera, Mio alpha and Microphthone

2.4 Experiment Design

Firstly, the experimenter installed the fixed wide-angle camera in advance, and arranged local area network to connect all measuring devices to the network and synchronize the clock. Half an hour before the start of the experiment, the experimenter wore a heart rate watch for the dispatcher, started the off-line recording function and started the wide-angle camera recording function. When all tasks were completed, the experimenter stoped recording the equipment and exported the data, so all the experimental data were collected [9].

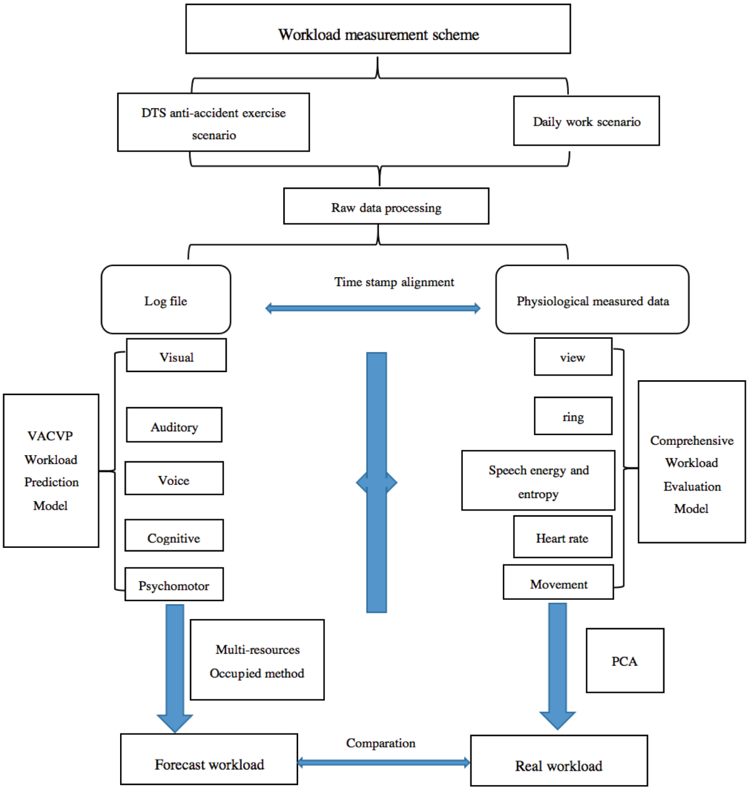

Workload measurement scheme
DTS anti-accident exercise scenario
Daily work scenario
Raw data processing
Log file
Time stamp alignment
Physiological measured data
Visual
Auditory
Voice
Cognitive
Psychomotor
VACVP Workload Prediction Model
view
ring
Speech energy and entropy
Heart rate
Movement
Comprehensive Workload Evaluation Model
Multi-resources Occupied method
PCA
Forecast workload
Comparation
Real workload

Fig. 4. The whole experimental process

The whole Experimental process as Fig. 4 shows. There were two models of data processing: the VACVP workload prediction model and the comprehensive workload evaluation model. In this experiment, the log file was recorded by the dispatcher based on the work plan and the physiological measurement data was obtained by the apparatuses. Two kinds of data were processed by time stamp alignment method to ensure data consistency. According to the log file, at a time point, the recorded task was decomposed into five resource channels and the workload was calculated based on the VACVP model. Meanwhile, according to the comprehensive evaluation model, the workload value is calculated by PCA to process physiological measurement data.

Finally, by comparing the values of predicted workload and actual workload to judge the accuracy and validity of VACVP workload prediction model.

2.5 Model

VACVP Workload Prediction Model. In VACVP, workload is composed of mental and physical loads [8], time and resource occupancy are taken as two main calculation indicators. From the perspective of multi-resource occupancy theory and information processing, brain load includes visual (V) auditory (A) and cognitive (C) in multi-resource occupancy theory, mainly in the information acquisition and information processing stage [7]. Physical load is related to human operation and movement, mainly includes psychomotor (P) and voice (V), focusing on the operational response stage [5, 7]. The whole process of task execution is described by information acquisition of VAV, information processing of C and motion P as response [7]. Each task

Table 1. VACVP rating scale

Resource access	Score	Description
Visual (V)	0	No vision
	1	Visual inspection, checking and processing
Auditory (A)	0	No auditory
	1	Auditory discrimination, feedback
Cognitive (C)	0	No cognitive
	1	Selection and signal recognition
	2	Symbol judgment and evaluation
	3	Assessment, judgment, memory (considering only one side)
	4	Assessment, judgment and memory (comprehensive multi considerations)
Voice (V)	0	No voice
	1	Simple answer
	2	Voice communication
Psychomotor (P)	0	No movement
	1	Discrete behavior (press buttons, keyboard input)
	2	Continuous behavior (walking)

type of the operator is assigned to five channels of VACVP in the multi-resource occupancy theory. Different channel occupancy weights are given according to the utilization of resource channels involved in task types. The proportion of channel resources occupied by each task was assessed by task analysis experts combined with VACVP rating scale, as shown in Table 1.

Furthermore, the workload of a task L in the same time period is weighted by the workload of each channel.

$$\mathrm{WL} = WL_V + WL_A + WL_C + WL_V + WL_P \tag{1}$$

Where: WL_v, WL_A, WL_C, WL_V, WL_P mean the workload of the channel Visual, Auditory, Cognitive, Voice and Psychomotor.

At a time point, if there are n tasks that operating simultaneously, the workload of serial and parallel tasks is added up by the workload of each task.

$$\mathrm{W}L_{total} = WL_{task1} + WL_{task2} + WL_{task3} + \ldots\ldots + WL_{taskn} \tag{2}$$

The corresponding relationship between VACVP evaluation and workload level is shown in Table 2.

Table 2. Mapping relationship between VACVP evaluation values and ranks

Weighted score	Workload level	Workload level description
0–1	1	Negligible workload
2–3	2	Low workload
4–6	3	Adequate residual capacity for additional tasks
7–9	4	Residual capacity is not enough to easily focus on additional tasks
10–12	5	Adequate attention cannot be given to additional tasks
13–16	6	Very little residual capacity, can only pay a little attention to additional tasks
17–20	7	Very little residual capacity, efforts can still ensure the normal conduct of affairs
21–30	8	Very high workloads result in almost no residual capacity, difficult to maintain the current level of effort
31–40	9	Extremely high workload. No residual capacity, hard to maintain the current level of effort
41–higher	10	Can not provide enough effort and can only abandon the task

The conversion formula for mapping the weighted score in Table 2 to the corresponding workload level is as follows.

$$y = 20 \times \left(\frac{1}{1 + 3^{-0.1x}} - 0.5 \right) \tag{3}$$

Where, x represents weighted score and y represents workload level.

Comprehensive Workload Evaluation Model. In this model, every measurement is distributed. Each measuring device synchronizes the internal clock through the LAN before each test, and then carries out the off-line measurement independently. Record the time stamp accurately to the millisecond level at each sampling point in the measurement process. After each recording is completed, the data files collected by the sensors are readed through the time axis synchronous calibration method [10], and the data is time aligned and integrated.

In this model, the dispatcher's behavior is analyzed by video processing. The motion is detected by the combination of skin color test and motion test. Firstly, the two frames f_n, f_{n-1} in the video sequence are subtracted, and the absolute value of the difference image is taken to get the corresponding difference image D_n.

$$\mathrm{D_n}(x, y) = |f_n(x, y) - f_{n-1}(x, y)| \tag{4}$$

Then, the threshold T binarization of D_n is processed, the connected row analysis is carried out to obtain the image contour R_n, which contains the complete moving object.

$$R_n(x, y) = \begin{cases} 255, & D_n(x, y) > T \\ 0, & else \end{cases} \tag{5}$$

At the same time, the skin color of the image is checked and processed to get the connected area with human skin color. The region correlation comparison is carried out on the R_n image to get the final target detection area and the dispatcher's motion index information. The motion results obtained from the above processing are shown in the following Fig. 5.

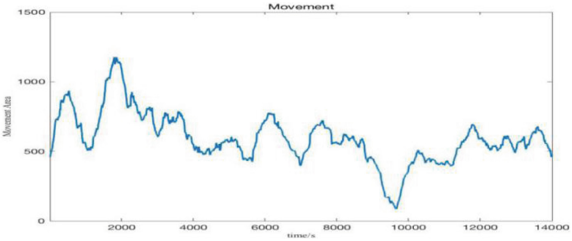


Fig. 5. Movement result

According to the different frequencies of human voice and ringtone, the upper and lower limit of cut-off frequency of voice signal passband is 100 Hz, 400 Hz, and the upper and lower limit of stop-band cut-off frequency is 50 Hz and 850 Hz respectively. The voice signal and ringtone are separated, the short-term energy and spectral entropy

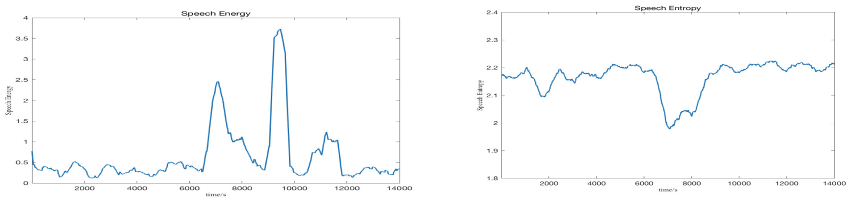


Fig. 6. Speech energy and speech entropy

of voice are extracted, so that the voice index information is obtained. The speech energy and speech entropy results obtained from the above processing are shown in the following Fig. 6.

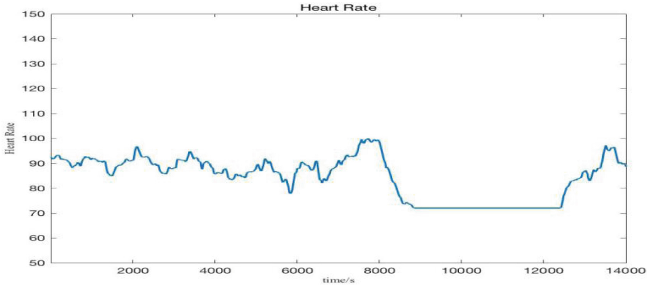


Fig. 7. Heart rate

The collected heart rate signals are filtered to eliminate the interference of external conditions, remove the extreme value of heart rate information interval, and retain the data of heart rate signals between 50 and 150. The heart rate results are shown in the Fig. 7.

After the above data processing, the comprehensive evaluation model can be mathematically set up. The workload of comprehensive workload evaluation model can be calculated by the following equation:

$$W = \beta_1 M + \beta_2 H + \beta_3 S \tag{6}$$

Where:

- Movement, Heart Rate, Speech energy are the results of physiological parameters stated above.
- $\beta_1, \beta_2, \beta_3$, are the weights to represent the contributions of movement, heart rate and speech energy. They are set by the algorithm called PCA (Principal Component Analysis).

In this paper, PCA is used to quantity the contributions of these factors (movement, heart rate, speech energy, speech entropy). The essence of PCA is the process of transforming the high-dimensional space into low-dimensional space, which makes the problem become more intuitionistic and simple [11]. Based on the above PCA processing, the β values of comprehensive evaluation model can be worked out as following equation:

$$W = 0.6902M + 0.1637H + 0.1379S \tag{7}$$

3 Result

3.1 Result of VACVP Workload Prediction Model

According to the VACVP model, the weighted scores of each channel in the same time period and the total forecasting workload level of the whole model can be obtained by

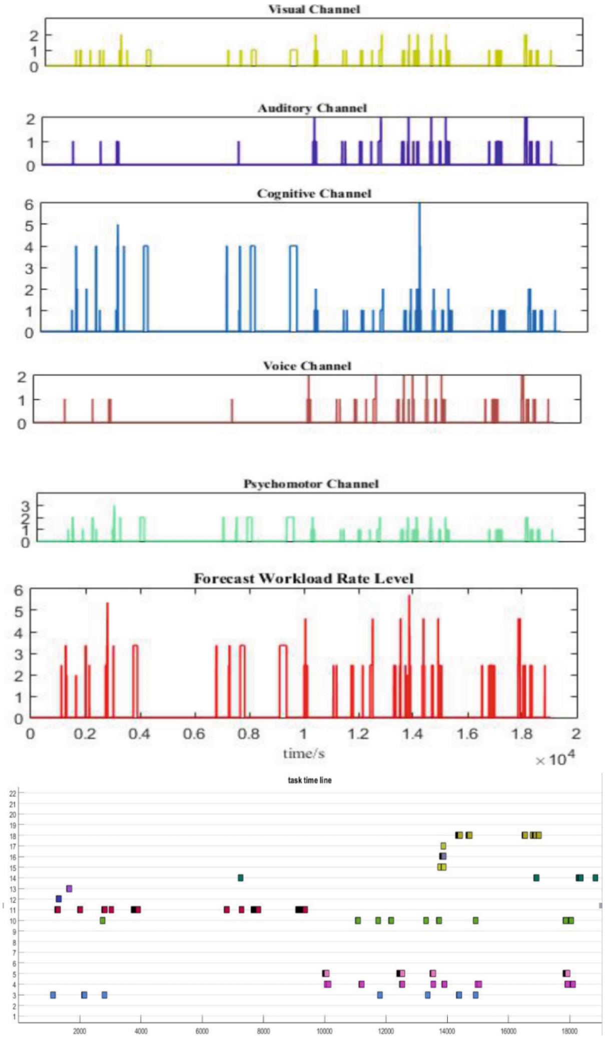


Fig. 8. Result of VACVP workload prediction model

processing the data of the task planning timeline. the results of VACVP model are as following.

The results in Fig. 8 clearly reflect the occupancy of every resource channel and the change of workload level during the execution of tasks. In particular, it is pointed out that the number in the longitudinal axis of the task time line graph represents the number of the current type of work, rather than the sum of the number of task types. If the task type occurs at a certain time, then a box appears at the number representing the task type, and the sum of the number of vertical blocks represents the total number of parallel task types.

3.2 Result of Comprehensive Evaluation Model

According to the comprehensive workload evaluation model, the results of actual workload level are as shown in Fig. 9.

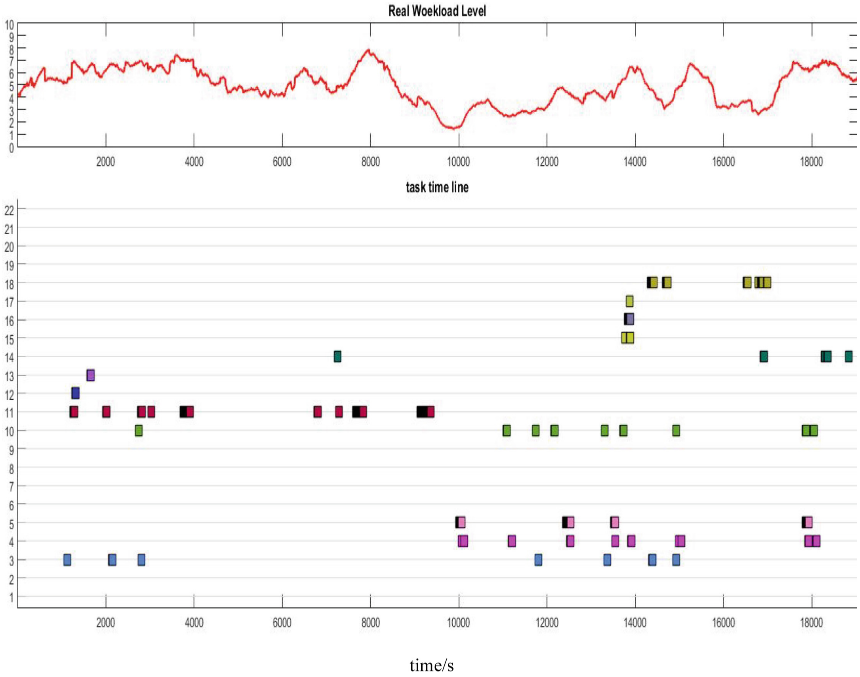


Fig. 9. Result of comprehensive workload evaluation model

In Fig. 9, the meanings of parameters in the task time line graph are the same as that in Fig. 8. According to Fig. 9, it can be concluded that the size of the workload is related to the complexity of a single task and the number of task types. When the complexity of a task is greater, its workload will increase. For example, when time = 3000 s and the task type is 11, the workload level is 5 (actually task 11 is indeed a complex task); when the number of task types is more, its workload will also increase. For example, when time = 14000 s, the sum number of tasks is 3, the workload level is 6.

3.3 Result Comparison

By showing the two workloads in the same time period, we can compare the values between the two models, the comparation results are as Fig. 10 shows.

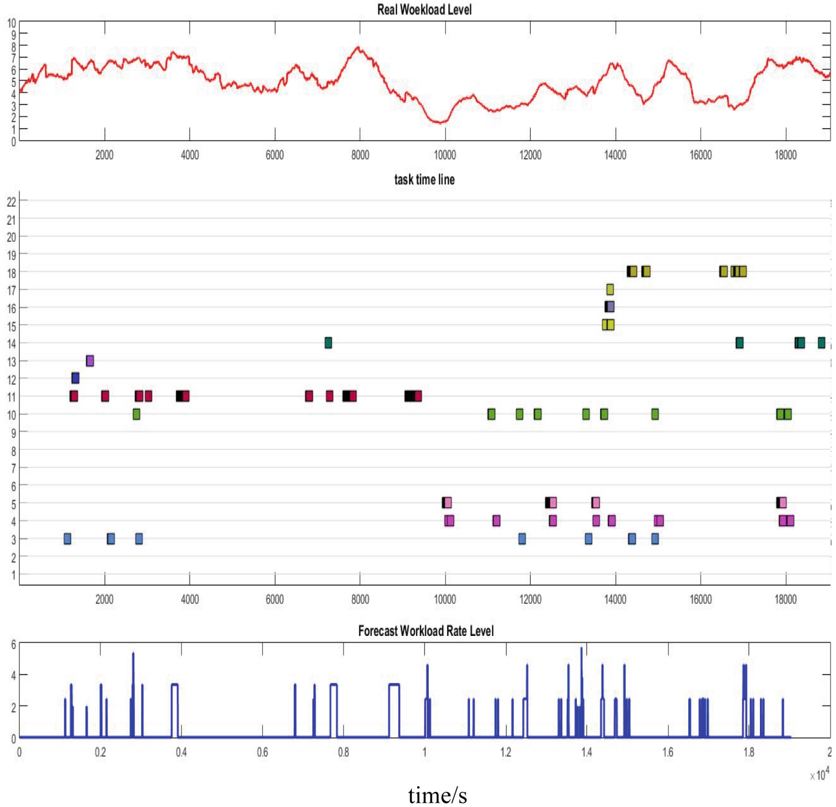


Fig. 10. Comparation of two models' results

From Fig. 10, we can see that the real workload and the forecast workload are positively correlated with the change of task type and total task amount in a certain period of time. This meanings the two workload measurement models are both effective and applicable to the actual working conditions.

By analyzing the changes of the two results data, it shows that the trend of workload forecast value is basically consistent with the actual value. While, the real value is slightly larger than the forecast workload value. Specific data results are shown in Table 3.

Table 3. Results comparison

Time/s	Actual workload level	Predicted workload level
5000	6.93	5.37
7000	4.78	3.36
9000	3.93	3.36
10000	1.60	2.45
11000	2.52	2.45
12000	4.20	2.45
13000	4.25	4.62
14000	3.63	2.45
15000	4.66	4.62
16000	2.62	2.45
17000	3.05	2.45

Table 3 shows the data of predicted and actual workload levels at specific time points, and the average error between predicted and actual workload levels is 0.6. This shows that within a certain error range, the predicted results of VACVP multi-resource occupancy model are correct and can accurately reflect the actual workload levels.

In the study of workload, the relationship between predicted value and actual value is validated. In Fig. 11, The trend of the two curves is basically consistent and relatively consistent, reflecting the positive correlation between them. At the same time, the accuracy and reliability of the VACVP workload prediction model are demonstrated.

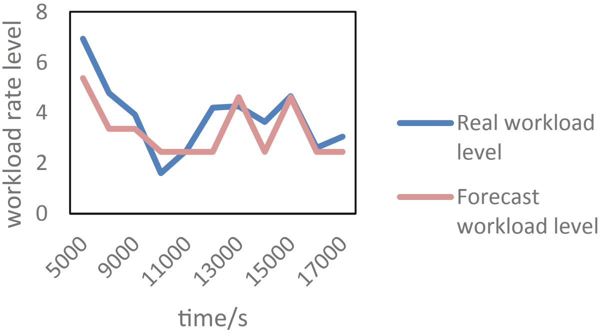


Fig. 11. Workload tendency comparison

4 Discussion

By comparing the two workload values, it is found that the predicted workload value is 0, while the actual workload value is not 0 in a certain period of time, it is because we neglect that the dispatcher has been monitoring. To solve this problem we can add a level to the definition of visual part: monitoring. There is monitoring in the whole process of duty, so there is always a non-zero value. For the other hand, due to the work plan is rough and can not be accurate to every moment, its values can be discrete, so in the process of prediction, there will be a situation that the predicted value is null between the two task types. In this solution, we can refine and improve the working log files to achieve time continuity.

5 Conclusion

In this paper, a multi-resource occupancy model VACVP is used to predict the workload of dispatchers in power grid for serial and parallel tasks. At the same time, a comprehensive evaluation model based on physiological parameters is constructed to validate and compare the prediction results. As shown from the experimental results there is a certain gap between the two kinds workload value and the average error can be maintained within 0.6 level. Furthermore, the trend of the two values in the same time period is basically consistent. These results reflect the validity and accuracy of the workload forecasting model.

The forecasting model proposed in this paper is generally used to predict the load in the actual working situation of the power grid. The accuracy and continuity of this method will be further improved and verified in future experiments.

References

1. Yerkes, R.M., Dodson, J.D.: The relation of strength of stimulus to rapidity of habit-formation. J. Comp. Neurol. Psychol. **18**, 459–482 (1908). https://doi.org/10.1002/cne.920180503
2. Hart, S.G.: NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Sage Publications (2006)
3. Kim, J.H., et al.: Measurement accuracy of heart rate and respiratory rate during graded exercise and sustained exercise in the heat using the Zephyr BioHarness. Int. J. Sports Med. **34**, 497–501 (2013)
4. Wilson, G.F.: An analysis of mental workload in pilots during flight using multiple psychophysiological measures. Int. J. Aviat. Psychol. **12**, 3–18 (2012)
5. Wickens, C.D., Yeh, Y.Y.: The dissociation between subjective workload and performance: a multiple resource approach. In: Proceedings of the Human Factors Society Annual Meeting, vol. 27, no. 3, pp. 244–248. SAGE Publications, Los Angeles (1983)
6. Mccracken, J.H., Aldrich, T.B.: Analysis of selected LHX mission functions: implications for operator workload and system automation goals: TNAASI 479-24-84. Fort Rucker: Anacapa Sciences (1984)
7. Wickens, C.D.: Muhiple resources and mental workload. Hum. Factors: J. Hum. Factors Ergon. Soc. **50**(3), 449–455 (2008)
8. Kohlmorgen, J., Dornhege, G., Braun, M., et al.: Improving human performance in a real operating environment through real-time mental workload detection. Toward Brain-Comput. Interfacing 409–422 (2007)
9. Distasi, L.L., Antoli, A., Gea, M., et al.: A neuroergonomic approach to evaluating mental workload in hypermedia interactions. Int. J. Ind. Ergon. **41**(3), 298–304 (2011)

10. Wang, Z., Fu, S.: An analysis of pilot's physiological reactions in different flight phases. In: Harris, D. (ed.) EPCE 2014. LNCS (LNAI), vol. 8532, pp. 94–103. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07515-0_10
11. Niu, Z., Qiu, X.: Facial expression recognition based on weighted principal component analysis and support vector machines. In: 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE). IEEE (2010)

Development of a Quantitative Evaluation Tool of Cognitive Workload in Field Studies Through Eye Tracking

Giovanni Pignoni(✉) and Sashidharan Komandur

Department of Design, Norwegian University of Science and Technology,
NTNU, 2802 Gjøvik, Norway
postmottak@ntnu.no
https://www.ntnu.edu/design

Abstract. Eye tracking is mainly employed as mean of tracking visual attention of an observer/operator. Still, eye tracking is also capable of recording a wider variety of data such as traces of mental workload. Pupil diameter have been validated as such measure. Most of the studies that have validated this are in laboratory conditions, where the perceived luminance (measured in candela per square meter) can be controlled. Luminance affects the pupil diameter as well; this means if the pupil diameter varies for an operator/observer in field conditions it cannot be accurately determined if the change in the pupil diameter is due to mental workload alone. Although there are some studies, which have attempted to simultaneously account for the contribution of the change in pupil diameter due to luminance and mental workload, not many have attempted to account for this in field conditions for safety-critical systems such as a helicopter or a maritime ship bridge. In this study as a first step, we define a method to measure luminance while tracking the gaze point. We will record eye-tracking data simultaneously recording the video feed of the field of view of the operator/observer. We will use the video feed to estimate the luminous flux from the point of view of the subject. We will be collecting this data from a helicopter pilot and his co-pilot during an actual operation (e.g. transportation of personnel and carrying a payload for an electrical power provider company in Norway or Sweden). We will also be collecting data from a navigator and his first officer in a high-speed marine craft of the Norwegian navy. We will also be collecting subjective data using paper-based tools such as NASA-TLX in addition to a conventional video recording of the scene of activity and handwritten notes of observation for validation purposes. We will also capture mental workload data from a few other objective sources such as heart rate variability (ECG). We expect to clearly define an approach to separately account for the effect of mental workload independent of the impact of changing light conditions in field situations for safety-critical systems. This includes a mathematical model that we innovate based on other mathematical models that are already available in the literature.

Keywords: Eye tracking · Pupillometry · Cognitive workload · Field study · Maritime usability

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 106–122, 2019.
https://doi.org/10.1007/978-3-030-22507-0_9

1 Introduction

Pupil dilation is an important metric for assessment of mental workload [15,21], especially in a safety critical system [8] such as a helicopter (or) ship's bridge or driving [40]. In these environments, the light condition fluctuates dramatically. Since changing light conditions can also impact pupil dilation, it is necessary to separate the effect of the mental workload from the effect of the changing light conditions to be able to utilise it reliably to evaluate the risk profile.

As discovered through the literature review, currently, there is no open and validated method to measure cognitive workload in a field condition trough pupillometry. The only commercially available method is proprietary [26] ad thus closed source, it disconnects the researcher from the ability to adapt the tools to specific research questions and have a deep understanding of the variables at play [17]. Moreover, no methods have been validated for use with low-cost eye trackers, which would enable affordable data gathering, including collaborative studies with multiple eye trackers.

The relation between cognitive workload and pupillary responses has been assessed as back as 1964 [15], here Hess and Poltmeasured changes in the pupil size of a subject during the resolution of "simple multiplication problems" and observed a link between the pupillary response and the difficulty level. Likewise, Kahneman et al. [21] investigated the correlation between task-evoked pupillary diameter and memory intensive tasks; reporting different pupillary responses from the learning and report/recollect phases as well as variations directly related to the task difficulty. These initials results were validated for a variety of "intensive cognitive tasks", including language, writing, listening, speech and the solution of mathematical problems [22]. The psycho-physiological studies on workload and the pupillary response are often limited by three main factors:

- Perceived luminance, as the variable with the more significant influence on the pupil diameter, it can mask the influence of cognitive workload. It is common to keep luminance as controlled conditions to isolate its effect.
- Real-time tracking is only possible tough a limited number of the reviewed ed methods. Online tracking of cognitive workload, [42] and [26] requires high-temporal-resolution, as well as the control of the environmental variables, including the estimate of a continually changing baseline value. Lack of such features limits the study to the evaluation of well defined/separated tasks.
- Open source. The only method that currently provides a solution to both the previous limitation is tied to patented technology, only limited documentation is available on the underlying method [26], and this makes it impossible for independent researchers to reuse, adapt and improve on such systems.

1.1 Research Questions

This paper is part of an ongoing, work-in-progress, research and will present only preliminary results to the following research questions:

- How to reliably measure luminance from the POV of a subject using a small calibrated video camera?
- How to calculate the baseline pupil size for a visual stimulus and use it to infer the cognitive state of the subject then?
- How to scale this method in field conditions (where luminance variate in an unpredictable manner)?

1.2 Planned Contributions

- Open source code, available in a public repository that can work on files generated by at least one common eye-tracker vendor.
- Thesis report and paper with validation data.

1.3 Existing Research

Cognitive Workload. Mental workload can be evaluated through a variety of methodologies [5,35]:

- Subjective-empirical measures of perceived effort as rated by the subject.
- Performance of the subject in a controlled task.
- Physiological indices of cognitive state.

Subjective reports such as questionnaires and multidimensional ratings (e.g. the NASA-TLX [14]) are indirect means of evaluation of the perceived workload. They are usually easy and cheap to administer but have several limitations [39]. As post-facto evaluations, relying on the personal impression and memory of the subject, they do not track a change over time and are therefore difficult to use for the identification of specific peaks on cognitive workload. The NASA-TLX questionnaire is a standardised assessment tool of cognitive workload. It employees a "multi-dimensional rating scale" measuring six parameters to give an estimate of the overall task workload: mental, physical, temporal, frustration, performance and effort. Like other forms of self-report, it doesn't record changes in cognitive load over time (multiple questionaries can be used to assist a complex task if divisible in subtasks). Other subjective workload measures are: Multiple Resources Questionnaire (MRQ) [2], Subjective Workload Assessment Technique (SWAT) [34], Overall Workload Level (OWL) [20] and Integrated Workload Scale (IWS) [31].

Performance-based measures of mental workload indirectly measure the cognitive state of a subject through the execution of a standardised task. Changes (speed, accuracy, response time) in the execution of the secondary task can be interpreted as a difference in cognitive/visual workload. The ISO defined Detection Response Task (DRT) [19] is an example of a performance-based method, Cegovnik et al. [4] used a tactile DRT to validate the use of a Tribe eye tracker to assess changes in the cognitive load of the subjects using oculography and pupillometry. The DTR estimates the cognitive load trough response rate and miss-rate of the response task: a stimulus is delivered through a vibrator attached to the subject in a random sequence; the test measures the response time (time needed to press a button attached to the steering wheel in a driving simulator). The use of this class of methodologies is to be planned considering the effect of the controls tasks on the main task as well as the low temporal resolution of the events that can be measured. Moreover, the relationship between cognitive workload and task performance is not linear and follows an inverted U-shape as defined in the Hebb-Yerkes-Dodson Law [4]. Both overload under-load can, therefore, result in decreased performances making the measure potentially unreliable if not paired with subjective data.

Physiological indices indirectly connect a measurable psychophysical parameter to an expected mental workload. Heart rate, respiratory rate, galvanic skin response, brain activity (EEG, fRMITCD), as well as eye activity [3] are parameters that over the years have been used to measure the mental workload. Modern eye tracking has an intrinsic advantage of being unobtrusive, and less impending that most of the other aforementioned techniques and could be a reliable instrument for over time monitoring of the mental workload [5]. Still, psychophysiological measures have several limitations [4] when applied in field conditions. In most of the studies mentioned beforehand, the experimental design included two or more tasks with different levels of difficulty indirectly estimating the workload required for each of the tasks. The physiological and neurological models employed in the psychophysiological methods have to be specially designed and trained to fit a particular task-evoked neural activity. It is, therefore, difficult to compare the results to a generalised measure of workload.

Blinks. Eyeblink duration and rate have been identified as an alternative metric for visual workload [1] and [33]. Unfortunately, this metric reliability is limited, as the blink rate can be influenced, in an opposite manner, by both the mental workload and visual workload.

Pupil. In the Handbook of Psychophysiology [3, p. 443], Cacioppo et al. defines the pupillary system as a "dually innervated organ". The pupil size is determined by the concurring action of a parasympathetically innervated constricting muscles and sympathetically innervated radial dilator muscles. The parasympathetic activity is dominant, responding to light reflexes, and determine the varying pupil size baseline, the sympathetic activity is instead connected to behavioural and stress contexts and can be used as a psychophysiological parameter of cognitive activity.

Task-Evoked Pupillary Response. Palinko et al. [30] estimated the driver's cognitive load from pupil size measurements finding that it the pupillary response correlates with the measured driving performances, and this as similar studies seems to confirm the reliability of pupillometry as a measure of cognitive workload. However, the analysis is limited to a simulated task with low variability between target luminance. Palinko et al. [30] introduced a pupillometric cognitive load measure for real-time cognitive load changes (every several seconds).

Light and Cognitive Load Effects on Pupil Diameter. Palinko et al. [29] follows up the previous study with a proof of concept of the possible separation between cognitive and ambient light components of pupil dilatation. The study was conducted using a driving simulator as a controlled environment for both an Aural Vigilance Task an Illumination Task (with different brightness targets) and a combined validation task. The study results show that it should be theoretically possible to model the psychophysical functions of the pupillary response over time to light stimuli and shows the measured trend over time. It also notes how the transitions bright/dark/bright are not equal as different muscle groups are involved in the contraction and dilation movements. The bright light reaction is quick "to protect the retina from overexposure", while the reaction to darkness is slower and gradual. The psychophysical function to predict an expected baseline pupil diameter should, therefore, take into account multiple parameters, current light level, previous light level, the rate of change, as well as age, and target. The study concludes that it is possible to discern the effects of luminance and cognitive load on pupil diameter and that the "proof of concept" predictor works in the limited experimental setting.

Discern Between Mentally and Visually Workload. Recarte et al. [32] validated the use of pupillary response as workload index in a field scenario ignoring the effect of illumination changes as the variable was impossible to control. The data they collected shows consistent results across the different driving task (no task, verbal task and visual task) collected during multiple driving session, to such a degree that cannot be explained by the sole different lighting conditions.

Unified Formula for Light-Adapted Pupil Size. Since the pupil diameter can be deconstructed as the result of multiple concurring factors, in order to correctly differentiate the cognitive workload from the pupillary light response, it can be useful to compute the expected pupil diameter for a given brightness condition and use the resulting value as a baseline value upon which calculate the cognitive-driven component of the measured pupil size. In a recent paper, Watson et al. [41] (NASA Ames Research Center and University of California) have reviewed seven different published psychophysical functions defining the relation between target luminance (cd/m^2) and expected pupil diameter. In the same paper, they also published a newly developed unified formula. The calculated baseline would work in the range of 2 to 8 mm, the reliability of the unified formula have to be tested to ensure that the little variability rage of the pupil size provoked by cognitive workload is preserved ($<\pm 1$ mm) [29]. The unified formula [41] is valid only for a light-adapted condition with stable illuminant and point of view (PoV) as it doesn't account for the adaptation state or the "pupillary unrest" (low-frequency random fluctuation in the range of 0.02 to 2.0 Hz and amplitude within ±0.25 mm).

Independent variables in the unified equation:

- Luminance.
- Age (The maximum pupil diameter, as well as total range, declines as the age grows).
- Field diameter (deg).
- Number of eyes stimulated, the final diameter is dependent on the number of eyes that are adapted to the light condition they defined the “Effective corneal flux density” (the variable controlling the effective pupil diameter) as dependent on the number of eyes (attenuated by a factor of 10 for one eye).

$$F = LaM(e)$$

(F) Flux density as the product of (L) luminance, (a) area, and (M(e)) monocular effect.

Wavelet Analysis. Marshall et al. [26] describes a technique to identify the origin of a recorded pupillary response that works independently from the target luminance. The procedure employs wavelet analysis to identify the dilation reflexes of the subject’s pupil. She explains how the reflexes can be differentiated as the pupil have different responses to light and psychosensorial stimulus. In a steady light, the pupil shows an irregular pulsation (light reflex) provoked by the interaction of the circular contracting muscles (agonist) and antagonist radial muscles act as the antagonist and are inhibited from dilating the pupil. A cognitive workload provokes a different waveform as both circular and radial muscles dilate the pupil creating a brief peak. This would imply that the cognitive workload is measured as the frequency and intensity of such events and not as a steady dilatation of the pupil (for the duration of the load), but it is unclear how this method would perform in a field condition, with highly variable ambient luminance.

An application of this technology is explained in a study conducted by Marshall for the US Navy. She applied the patented metric of Index of Cognitive Activity (ICA) [27] thanks to a networked system that is set up to record the cognitive workload for multiple team members during a collaborative task. The study assessed the performance of a three-person team in a simulation system, and the effort to overcome mission-related problems. A similar study, a collaboration between NASA Ames and EyeTracking, Inc. used the ICA and eye metrics to detect the difference between low and high fatigue states [28].

Machine Learning for Pupillometry. Wierda et al. [42] and the related work of Ferscha et al. [9] represent a different approach to the problem of the indirect assessment of mental workload. As the response time of the pupil to a mental workload event is too slow (several seconds) to be used as a real-time measure, it can be used directly only as an average over time. This makes it suitable to evaluate lengthy tasks that have a reasonably constant load in cognitive workload (at least several seconds). These two studies show a proof of concept of how to obtain an high-temporal-resolution (c.a. 10 Hz) tracking of the cognitive processes through deconvolution. The aim of real-time cognitive workload measurement gains value in the context of the implementation of a real-time feedback loop in the interaction design of a system (e.g. a system able to respond to different cognitive states of the user).

Wierda et al. [42] fixed the distribution of "attention impulses" every 100 ms defining the output's temporal resolution. Employing a model of the "Task-evoked pupil impulse response," it reconstructs the intensity of the attention impulse that provoked the measured pupillary response. Ferscha et al. [9] further developed the concept through machine learning for better performances without the need of a fixed temporal resolution of the cognitive impulses. To reduce the effect of incident light Ferscha et al. [9] used the average illumination in the subject's field of view analysing the eye tracker camera stream. The technique they used is possibly still insufficient to adapt the technology to a field study with a highly variable illumination. In the described implementation a luminance change more significant than the set threshold would trigger a suspension of the tracking, this state is then maintained until the condition is stable again and a new baseline can be calculated. A similar solution was implemented to filter out blinks. The dynamic baseline is computed through a series of threshold and doesn't adjust for small changes in target luminance.

Illuminance Measurement Using a Digital Camera. Luminance as a measure needed to dynamically estimate the pupil size of a subject, candela per square meter cd/m^2, is the quantity of light radiating from a source. An illuminance meter is an expensive and bulky device, in more than one instance this has resulted in attempts to use a camera as cheaper and more flexible alternative [16] and [43].

A digital sensor is at its core an array of Illuminance sensors. Each pixel measures the number of photons hitting the photoelectric surface. The presence of a Bayer filter for colour photography makes it so that to reconstruct the information form the entire visible spectrum multiple pixels have to be analysed at the same time. Each pixel in the final image has reconstructed values from the neighbouring pixels for all the RGB channels and in itself would be sufficient to reconstruct the illuminance of the scene, in order to reduce noise and gain reliability multiple pixels should be used, the number of pixels used is effectively the field of view of the instrument [16]. The formula proposed by Hiscocks [16] has been optimised for a DSLR camera, not all the parameters are accessible when using an embedded digital video camera.

Parameters:

- Pixel value (0–255 for an 8bit monochrome image).
- Shutter Speed (In a video camera, this is limited by the frame rate, e.g. 1/30 s for a 30 fps camera) and aperture or focal ratio.
- Iso or film speed.
- Camera Constant (The calibration constant for a specific camera model that has to be determined with a known instrument).

Wuller [43] suggest a different model, reversing the colour processing of the camera, first from gamma-compressed RGB to linear RGB and then from linear RGB to CIE XYZ, extrapolating then $y(\lambda)$ as the relative luminance (luminance as defined by the luminosity function, reproducing the spectral luminous efficiency of the human eye). The author notes that to access the linear response of an image sensor the correct inverse gamma has to be applied and that this could deviate from the standard 2.2 (sRGB), the relative luminance can then be converted to luminance through a linear relation specific for a particular sensor/camera settings combination.

Maritime Usability and SA. Endsley et al. [6] defines situational awareness (SA) as "the perception of the elements in the environment within a volume of space and time, the comprehension of their meaning and the projection of their status in the near future". Low SA has been found to be one of the primary sources of human error in safety-critical systems [36]. Real-time monitoring of SA seems possible through the analysis of the subject visual attention aided by a variety of eye tracking data such as:

- Fixation duration: length of fixations (e.g. time spent on a single target without movement).
- Fixation rate: average number of fixations in a unit of time.
- Dwell time: the sum of all the fixation time in a single area of interest.
- Saccadic main sequence: the relation between the saccadic duration and magnitude and between peak velocity (PV) and magnitude [5], as both PV and duration increase with the magnitude.
    - Saccadic duration: the period between two positions of the fovea.
    - Saccadic magnitude: the magnitude of the saccadic movement (angle).
    - Peak saccadic velocity: highest velocity reached during saccades deg/sec.

Use of SAGAT. The Situation Awareness Global Assessment Technique (SAGAT), is a global tool developed to assess SA [7]. "A simulation employing a system of interest is frozen at randomly selected times, and operators are queried as to their perceptions of the situation at that time. The system displays are blanked, and the simulation is suspended while subjects quickly answer questions about their current perceptions of the situation. As a global measure, SAGAT includes queries about all operator SA requirements, including Level 1 (perception of data), Level 2 (comprehension of the meaning) and Level 3 (projection of the near future) components. This includes a consideration of system functioning and status, as well as relevant features of the external environment. SAGAT queries allow for detailed information about subject SA to be collected on an element by element basis that can be evaluated against reality, thus providing an objective assessment of operator SA."

Ikuma et al. [18] compared different standard human factors measurement tools: workload ratings (SWAT and NASA-TLX) and Situation Awareness Global Assessment Technique (SAGAT). Eye tracking was also used to analyse the gaze path of the participants during the simulation, "the percentage of time spent looking at different areas of the screen during steady-state periods differed among workload levels". This study only looks at a small number of areas of interest (AOIs) on the interface, to infer the visual attention of the subjects for different areas of the interface. The usability of on-board interfaces on High-Speed Craft (HSC) has been assessed through the application of eye tracking technology [11]. The cognitive workload and SA of the crew of a military HSC in littoral waters were selected as of interest because of the combination of high-speed navigation and the need to navigate outside established routes. With particular interest on the role of the navigator [10] and its use of the onboard interface "Route Monitor Window". Hareide et al. [11] collected data from both field and simulator activities using the Tobii Pro Glasses 2 Eye Tracker. The methodology of the study tried to account for the difference in the environment and datasets between the simulator and field conditions. The Author followed up a mid-life update of the interface [12,13], with further validation of the redesigned interface for the primary objective of increased navigator attention dedicated to the "outside" Area of interest opposed to the various interfaces. Eye trackers were in this case used as to indirectly evaluate situational awareness of the navigator through quantisation of the time spent on the interface rather than observing the environment. Hareide et al. [12] apply the concept of dwell time, look-backs and Backtracks to the analysis of AOIs:

- Look-backs (returns, refixation) are saccades landing in an AOIs already visited. The analysis of a look-back can point to a variety of concurring factors: memory failure, confusion on the function of a command/element, the difficulty of content understanding and intrinsic importance of the information present in an AOI.
- Backtracks are calculated on the specific sequence of saccades and are a sudden (inverted gaze direction) rapid eye movement back to a just visited AOI. Confusion or uncertainty, changes in goals, a mismatch between the users' expectation and interface layout.

The author shows how eye tracking data can be used to guide the development of a GUI through the analysis of areas of interest and gaze behaviour, but also notes several limitations in the use of the eye tracker that need to be considered not to influence the behaviour of the user group. This includes the thickness of the eyepiece frame, creating a visible "frame of vision", unwanted reflection and glares on the protective glass, difficulty using the binoculars in conjunction with the trackers and unfavourable lighting conditions.

2 Methods

2.1 Measure of Luminance from POV

This is a work in progress. his research starts with the development of the necessary tools. Even though it is theoretically possible to use a video camera as luminance meter, the reliability and accuracy of this technique will depend significantly on the software and hardware. The calibration and validation of the equipment will be done with a know good instrument in the Norwegian Colour and Visual Computing Laboratory. A colour checker will be measured with both the Konica Minolta CS-2000 [25] spectroradiometer and the World Camera mounted on the Pupil lab eye tracker (Pupil Pro). The data can then be used to calibrate the World Camera as a rudimentary luminance meter. A colour checker illuminated by a diffuse light at a variable intensity will be measured through both the Pupil Pro and the spectroradiometer. Different combinations of illumination and exposure settings on the software are required to model the sensor response.

The Pupil lab [23] software in his current version (1.10.20) does not dynamically save the exposure settings during the recording. Libuvc [38] (cross-platform library for USB video devices) is used to receive the video stream and communicate with the two cameras. Libuvc supports either getting or setting the exposure value and should allow retrieving the current exposure data during the recording. The lack of support of these functionalities in the Pupil lab software makes it impossible to use automatic exposure as the calibration values would change during the recording in an unpredictable manner. Using a fixed manual exposure is possible but severely limits the maximum dynamic range of the light meter.

Alternatives to the use of the camera, to simulate a scenario in which the aforementioned limitation does not apply, which would be easily reachable with some interest from the developers, is to use an external light meter mounted on the eye-tracker. This would provide a measurement that is not bound to the limited dynamic range of the camera with the drawback of having two disconnected data streams.

Instrumentation. The Konica Minolta CS-2000 spectroradiometer [25] is a high precision polychromatortype spectroradiometer, it will allow measurement on a vast range of luminance (0.3 to 500,000 cd/m^2) with a ±2% accuracy. The Pupil Pro [23] and [24] World Camera is mounted just above the subject eye, facing outward. The camera offers different combinations of resolution and framerate 1920 × 1080 @30 fps, 1280 × 720 @60 fps, 640 × 480 @120 fps covering a FoV of 60 or 100° diagonally (depending on the lens).

2.2 Calculate the Baseline Pupil Size

The measure of luminance from the POV will be connected to the unified formula for light-adapted pupil size developed by [41]. Two elements will need validation:

- the accuracy and precision of the unified formula.
- the different possible methods to convert the input from the camera to the correct input values for the formula.

The [41] unified formula is based on a standard procedure involving a defined stimulus: the observer is shown a bright circle on a dark background. The size (degrees of field of view) and luminance (cd/m^2) of the circle determine the corneal flux density (i.e. the product of luminance and subtended area) as defined by [37]: $D = 7.75 - 5.75[(F/846)0.41/((F/846)0.41 + 2)]$ “where D is the pupil diameter (mm), and F is the corneal flux density (cdm-2deg2)”.

The model implies that at its core the pupil control mechanism reacts as a ‘flux integrator’, following an S-shaped curve.

An image from the camera has to be used to evaluate the flux density or to indirectly convert the image into a corresponding standardised stimulus (i.e. a circle on a dark background). The most promising approach is to consider the average luminance on the camera sensor (or external light sensor) as equal to the luminance of a standardised stimulus (bright circle) as wide as the entire field of view (FoV) (120–190°).

This assumption ties the precision of the calculated pupil size to how well the camera FoV matches the user FoV). To test the quality of the model NTNU students and staff (age from 20 to 50) will be recruited for validation in a controlled environment.

Validation. The procedure will refer to the methods used by Palinko et al. [29]. It will be divided into three parts:

- No-load - variable light-adapted state. The participants will be sitting in a dark room looking at a selection of projected images; the sitting position will be adjusted to maintain a constant field of view and distance from the screen. The projected images will include standardised stimuli as well as more complex images (e.g. outdoor naturalistic scenery). Each image will be represented for several seconds to let the pupil reach an adapted state (c.a. 15 s). Each image will include a focal point ad the participants will be asked to stare at the focal point.
- Load - static light-adapted state. With a constant ambient luminance (e.g. grey image projected), the participants will be asked to perform an Aural Vigilance Task (AVT), as in [29]. The task involves listening to a voice counting from 1 to 18, repeated multiple times. Every 6th number (6, 12, and 18) might contain errors (i.e. another number is replaced to the correct sequence). The participants would have to perform an action such as pressing a button when they detect an error. The task should induce an increased cognitive workload near every 6th number. The location of the error should be randomly selected for each session.
- Load - variable light-adapted state.
  The same AVT task is repeated but with variable standardised images being projected.

2.3 Measure of Luminance from POV

Subjects. The subjects for the final session will be recruited as cadets of the Royal Norwegian Naval Academy (RNoNA) and will require access to the training vessel (Kvarven). The crew of a training vessel includes navigator, assistant, helmsman and training instructor. During a study session, the vessel would also include one to three researchers to set up and record the data. Depending on the availability of multiple eye trackers both the navigator and helmsman could participate in the experiment for each session.

Procedure. The setup for each session will include:

- Introduction to the research and signature of the informed consent.
- Application of the eye tracker (glasses and recording device).
- Calibration of the eye tracker.
- Reference measurements of ambient illumination.
- The debriefing will include a short interview and the NASA-TLX questionnaire as a further reference of the cognitive workload.

Two researchers should be present on board at any time. Between each session, up to 30 min will be required for cleaning of the instrumentation, download of the test data and recharge of the various batteries.

The task aims to highlight different levels of cognitive workload. The navigation task should be repeatable Fig. 1, and it should last less than one hour (not including the setup and debriefing) and should include a mix of low and high workload for the subjects: E.g. Steady navigation - change of course - steady navigation.

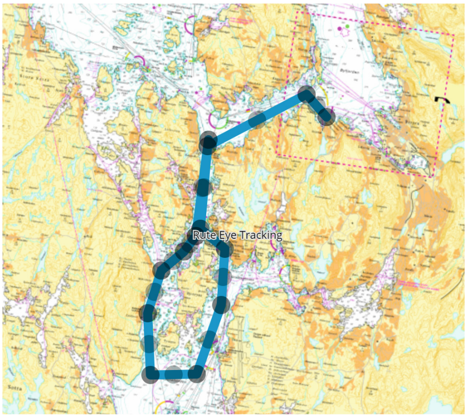


Fig. 1. The course suggested by Odd Sveinung Hareid from Laksevåg (Bergen)

3 Expected Results

3.1 Proof of Concept

This is a work in progress, at the time of writing, the research is still in the initial exploratory phase and should be completed by the end of May 2019. The literature review helped in defining a path to follow in order to develop the necessary tools (the pupil baseline calculation), but there are inherent risks and challenges in the generalisation of findings that were initially only meant for controlled condition/laboratory study. A series of non validated tests is being carried out to determine:

- weather the camera-based luminance meter works well enough for a reliable field application (limited dynamic range of the camera, limited bit depth, the difference in the camera FOV compared to the subject FOV).
- Weather pupil baseline calculation is precise to such a degree not to mask the cognitive workload.
- Weather the chosen eye tracker can operate in field conditions.

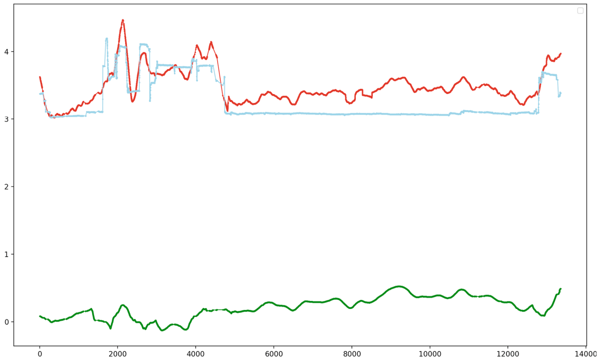


Fig. 2. This sample data output was recorded from a user sitting in front of a laptop in a dark room. The green line is the pupil size (mm) as measured by the eye tracker; the red line is the baseline as calculated from the video data and the blue line is the difference between the two. The blue will ultimately represent the cognitive workload. In this sample, the middle "steady" part has been measured on a subject performing an IQ test. The horizontal axe, time, is expressed in video frames at 30 fps. (Color figure online)

A series of artefacts have been identified in the sample data collected:

- Pupillary overshoot, the model of the pupil size is specific to adapted state and doesn't account for the pupil natural overshoot that can be observed when a rapid change in luminance occurs Fig. 2.
- Pupillary unrest, in the form of low-frequency random fluctuation in the range of 0.02 to 2.0 Hz and amplitude within ±0.25 mm Figs. 2 and 3.
- Incorrect measured pupil range; the pupil size is calculated from the video image in pixels to an estimated mm by the 3d model. different calibration of the pupil camera (distance from the eye) can bring the measured range outside the unified formula unified formula [41] range (c.a. more than 2 mm and less than 8 mm).

- Luminance outside the camera dynamic range, this is visible in the second plot, in this case, an outdoor recording ended inside a building, the completely black image from the video produces an unreliable luminance reading Fig. 3.

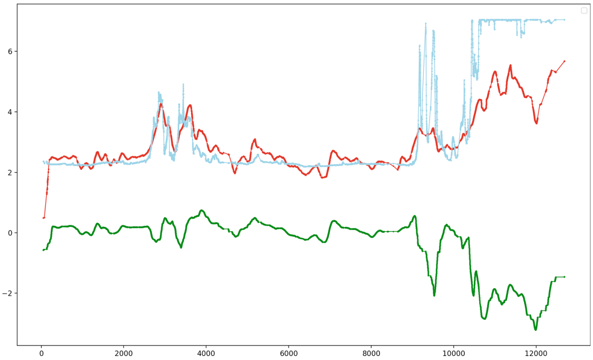


Fig. 3. This outdoor session shows the limitations of the camera if used with a fixed exposure.

During the test multiple data sources will be combined including: Heart rate variability, POV camera, eye tracking camera all will be used to substantiate the measure of cognitive workload. Dates and travel The number of sessions depends on the availability of cadets, five to ten subjects would be a good result. Given the nature of the experiment weather and ambient illumination conditions should be kept constant within a reasonable range. This could require the spread of the study over multiple days.

To account for the limitations of the eye tracker, it would be advisable to plan a portion of the sessions after dawn, (lower the contrast between the user interface inside the cabin and the outside).

4 Conclusion

The development of the necessary tools is in progress and will hopefully end as a refined proof of concept and validation of the method with the intent of attracting the interest of developers to consolidate the application. The validation that will be attempted as part of the research will be by no mean be exhaustive, it is expected that the interest surrounding the measure of the cognitive workload will result in a variety of experiments on the topic, to further explore the benefits and limitations of the developed methods.

The choice of a camera as a luminance meter could severely limit the accuracy of the method but would allow a variety of head-mounted eye trackers to be used for cognitive workload studies without the need of any extra hardware.

References

1. Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., Montanari, R.: Driver workload and eye blink duration. Transp. Res. Part F: Traffic Psychol. Behav. **14**(3), 199–208 (2011). https://doi.org/10.1016/j.trf.2010.12.001. http://linkinghub.elsevier.com/retrieve/pii/S136984781000094X
2. Boles, D.B., Adair, L.P.: Validity of the multiple resources questionnaire (MRQ). Proc. Hum. Factors Ergon. Soc. Annu. Meet. **45**(25), 1795–1799 (2001). https://doi.org/10.1177/154193120104502508
3. Cacioppo, J.T., Tassinary, L.G., Berntson, G. (eds.): Handbook of Psychophysiology, 3rd edn. Cambridge University Press, Cambridge (2007). https://doi.org/10.1017/CBO9780511546396. http://ebooks.cambridge.org/ref/id/CBO9780511546396
4. Cegovnik, T., Stojmenova, K., Jakus, G., Sodnik, J.: An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers. Appl. Ergon. **68**, 1–11 (2018). https://doi.org/10.1016/j.apergo.2017.10.011. https://linkinghub.elsevier.com/retrieve/pii/S0003687017302326
5. Di Stasi, L., Marchitto, M., Antolí, A., Cañas, J.: Saccadic peak velocity as an alternative index of operator attention: a short review. Rev. Européenne de Psychol. Appliquée/Eur. Rev. Appl. Psychol. **63**(6), 335–343 (2013). https://doi.org/10.1016/j.erap.2013.09.001. https://linkinghub.elsevier.com/retrieve/pii/S1162908813000741
6. Endsley, M.R.: Design and evaluation for situation awareness enhancement. Proc. Hum. Factors Soc. Ann. Meet. **32**(2), 97–101 (1988). https://doi.org/10.1177/154193128803200221
7. Endsley, M.R.: Direct measurement of situation awareness: validity and use of SAGAT. In: Situation Awareness Analysis and Measurement, pp. 147–173. Lawrence Erlbaum Associates Publishers, Mahwah (2000)
8. Engström, J., Johansson, E., Östlund, J.: Effects of visual and cognitive load in real and simulated motorway driving. Transp. Res. Part F: Traffic Psychol. Behav. **8**(2), 97–120 (2005). https://doi.org/10.1016/j.trf.2005.04.012. https://www.sciencedirect.com/science/article/pii/S1369847805000185
9. Ferscha, A., Gollan, B.: Modeling pupil dilation as online input for estimation of cognitive load in non-laboratory attention-aware systems. In: Eighth International Conference on Advanced Cognitive Technologies and Applications, COGNITIVE 2016, Rome, Italy, 20–24 March 2016 (2016)
10. Gould, K.S., Røed, B.K., Saus, E.R., Koefoed, V.F., Bridger, R.S., Moen, B.E.: Effects of navigation method on workload and performance in simulated high-speed ship navigation. Appl. Ergon. **40**(1), 103–114 (2009). https://doi.org/10.1016/j.apergo.2008.01.001. http://linkinghub.elsevier.com/retrieve/pii/S0003687008000094
11. Hareide, O.S., Mjelde, F.V., Glomsvoll, O., Ostnes, R.: Developing a high-speed craft route monitor window. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2017. LNCS (LNAI), vol. 10285, pp. 461–473. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_33
12. Hareide, O.S., Ostnes, R.: Maritime usability study by analysing eye tracking data. J. Navig. **70**(05), 927–943 (2017). https://doi.org/10.1017/S0373463317000182
13. Hareide, O.S., Ostnes, R.: Validation of a maritime usability study with eye tracking data. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2018. LNCS (LNAI), vol. 10916, pp. 273–292. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91467-1_22

14. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. **52**, 139–183 (1988). https://doi.org/10.1016/S0166-4115(08)62386-9. http://linkinghub.elsevier.com/retrieve/pii/S0166411508623869
15. Hess, E.H., Polt, J.M.: Pupil size in relation to mental activity during simple problem-solving. Science **143**(3611), 1190–1192 (1964). https://doi.org/10.1126/science.143.3611.1190
16. Hiscocks, P.D., Eng, P.: Measuring luminance with a digital camera, p. 27. Syscomp Electronic Design Limited (2014)
17. Holmqvist, K., Andersson, R.: Eye tracking: a comprehensive guide to methods and measures (2017). OCLC number: 1057373387
18. Ikuma, L.H., Harvey, C., Taylor, C.F., Handal, C.: A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. J. Loss Prev. Process. Ind. **32**, 454–465 (2014). https://doi.org/10.1016/j.jlp.2014.11.001. https://linkinghub.elsevier.com/retrieve/pii/S095042301400179X
19. ISO: Road vehicles – transport information and control systems – detection-response task (DRT) for assessing attentional effects of cognitive load in driving. ISO 17488:2016 (2016)
20. Jung, H.S., Jung, H.S.: Establishment of overall workload assessment technique for various tasks and workplaces. Int. J. Ind. Ergon. **28**(6), 341–353 (2001). https://doi.org/10.1016/S0169-8141(01)00040-3. http://linkinghub.elsevier.com/retrieve/pii/S0169814101000403
21. Kahneman, D., Beatty, J.: Pupil diameter and load on memory. Science **154**(3756), 1583–1585 (1966). https://doi.org/10.1126/science.154.3756.1583
22. Kahneman, D., Beatty, J., Pollack, I.: Perceptual deficit during a mental task. Science **157**(3785), 218–219 (1967). https://doi.org/10.1126/science.157.3785.218
23. Kassner, M., Patera, W., Bulling, A.: Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In: Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2014 Adjunct, pp. 1151–1160. ACM, New York (2014). https://doi.org/10.1145/2638728.2641695
24. Kassner, M., Patera, W., Bulling, A.: Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction, April 2014. http://arxiv.org/abs/1405.0006
25. Konica Minolta Sensing Singapore Pte Ltd.: CS-2000 SPECTRORADIOMETER. Konica Minolta Sensing Singapore Pte Ltd., March 2018. http://sensing.konicaminolta.asia/products/cs-2000-spectroradiometer/
26. Marshall, S.P.: Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity, March 1999. https://patents.google.com/patent/US6090051A/en
27. Marshall, S.P.: What the eyes reveal: measuring the cognitive workload of teams. In: Duffy, V.G. (ed.) ICDHM 2009. LNCS, vol. 5620, pp. 265–274. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02809-0_29
28. Marshall, S.: The index of cognitive activity: measuring cognitive workload. In: Proceedings of the IEEE 7th Conference on Human Factors and Power Plants, pp. 7:5–7:9. IEEE, Scottsdale (2002). https://doi.org/10.1109/HFPP.2002.1042860

29. Palinko, O., Kun, A.: Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. In: Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: Driving Assessment, pp. 329–336. University of Iowa, Olympic Valley-Lake Tahoe (2011). https://doi.org/10.17077/drivingassessment.1416, http://ir.uiowa.edu/drivingassessment/2011/papers/48
30. Palinko, O., Kun, A.L., Shyrokov, A., Heeman, P.: Estimating cognitive load using remote eye tracking in a driving simulator. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ETRA 2010, p. 141. ACM Press, Austin (2010). https://doi.org/10.1145/1743666.1743701
31. Pickup, L., Wilson, J.R., Norris, B.J., Mitchell, L., Morrisroe, G.: The integrated workload scale (IWS): a new self-report tool to assess railway signaller workload. Appl. Ergon. **36**(6), 681–693 (2005). https://doi.org/10.1016/j.apergo.2005.05.004. http://linkinghub.elsevier.com/retrieve/pii/S0003687005000918
32. Recarte, M.A., Nunes, L.M.: Effects of verbal and spatial-imagery tasks on eye fixations while driving. J. Exp. Psychol. Appl. **6**(1), 31–43 (2000). https://doi.org/10.1037/1076-898X.6.1.31
33. Recarte, M.A., Perez, E., Conchillo, A., Nunes, L.M.: Mental workload and visual impairment: differences between pupil, blink, and subjective rating. Span. J. Psychol. **11**(2), 374–385 (2008). https://doi.org/10.1017/S1138741600004406
34. Rubio, S., Diaz, E., Martin, J., Puente, J.M.: Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. Appl. Psychol. **53**(1), 61–86 (2004). https://doi.org/10.1111/j.1464-0597.2004.00161.x
35. Rusnock, C.F., Borghetti, B.J.: Workload profiles: a continuous measure of mental workload. Int. J. Ind. Ergon. **63**, 49–64 (2018). https://doi.org/10.1016/j.ergon.2016.09.003. https://linkinghub.elsevier.com/retrieve/pii/S0169814116301287
36. Sharma, C., Bhavsar, P., Srinivasan, B., Srinivasan, R.: Eye gaze movement studies of control room operators: a novel approach to improve process safety. Comput. Chem. Eng. **85**, 43–57 (2016). https://doi.org/10.1016/j.compchemeng.2015.09.012. https://linkinghub.elsevier.com/retrieve/pii/S0098135415003075
37. Stanley, P.: The effect of field of view size on steady-state pupil diameter. Ophthalmic Physiol. Opt. **15**(6), 601–603 (1995). https://doi.org/10.1016/0275-5408(94)00019-V. http://linkinghub.elsevier.com/retrieve/pii/027554089400019V
38. Tossell, K.: ktossell/libuvc (2017). https://github.com/ktossell/libuvc
39. Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics **39**(3), 358–381 (1996). https://doi.org/10.1080/00140139608964470
40. Voogt, A.D.: Helicopter accidents at night. Aviat. Psychol. Appl. Hum. Factors **1**(2), 99–102 (2011). https://doi.org/10.1027/2192-0923/a000013
41. Watson, A.B., Yellott, J.I.: A unified formula for light-adapted pupil size. J. Vis. **12**(10), 12–12 (2012). https://doi.org/10.1167/12.10.12
42. Wierda, S.M., van Rijn, H., Taatgen, N.A., Martens, S.: Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. Proc. Nat. Acad. Sci. **109**(22), 8456–8460 (2012). https://doi.org/10.1073/pnas.1201858109
43. Wuller, D., Gabele, H.: The usage of digital cameras as luminance meters, San Jose, CA, USA, p. 65020U, February 2007. https://doi.org/10.1117/12.703205

Towards a Mixed Reality Platform for Applied Cognitive Load Evaluation

Maurice van Beurden(✉) and Linsey Roijendijk

TNO Human Factors, Soesterberg, The Netherlands
maurice.vanbeurden@tno.nl

Abstract. Physical load experienced by dismounted military has already been studied extensively. Few studies, however, have focused on the cognitive load of dismounted soldiers during military operations in the field. In this project, a cognitive task analysis (CTA) was performed to study the cognitive workload of a troop commander during troop hasty attack trainings by the Royal Netherlands Marine Corps. Observations, interviews and questionnaires were used to study the cognitive load. The experienced cognitive and physical load was rated as high and the various phases of the attack contributed to this load differently. The CTA revealed that the cognitive tasks during the hasty attack were performed while being physically active or directly after heavy physical effort. This means that cognitive task performance should not be studied in isolation and the presence of physical activity may be an important factor moderating cognitive performance. The CTA further showed that situational awareness, task switching, and communication are important cognitive skills used by the troop commander. These findings are used to develop a mixed reality platform that can be used to investigate the effects of new technological innovations on cognitive performance under conditions mimicking real-life situations, while controlling for potential confounding variables.

Keywords: Cognitive workload · Cognitive task analysis · Simulation · Dismounted soldiers

1 Introduction

Dismounted soldiers are confronted with complex and physically demanding tasks. The expectation is that future soldiers will be equipped with advanced technologies aimed to support the execution of their tasks. The large amount of equipment will not only influence physical performance but may also impact cognitive performance as these new technologies are not always attuned to military operations. Physical load experienced by dismounted soldiers has already been studied extensively and is often measured using a military obstacle course. However, to investigate and understand the effects of new technologies (e.g., battlefield systems, augmented technology) on cognitive task performance in the military domain, no such framework exist. Current cognitive test batteries for soldiers particularly aim to predict future performance [1] or aim to identify individuals with Traumatic Brain Injury [2]. Moreover, only a few studies have focused on the cognitive load of soldiers during military operations in the

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 123–136, 2019.
https://doi.org/10.1007/978-3-030-22507-0_10

field as compared to research among pilots [3, 4]. Nevertheless, studying cognitive load among dismounted soldiers is important to identify how new (information) technology facilitate or impede cognitive performance.

In the current project, the aim was to gain insight into the tasks and cognitive load of a troop commander during a troop hasty attack with the purpose to implement a mixed reality platform that will be used to understand the impact of new technology on soldier's cognitive performance. The performance during a hasty attack was studied as such an attack involves the fundamental basics of infantry actions in conventional warfare and is also known to be challenging due to various uncertainties (e.g., unknown threat location, limited terrain knowledge) caused by a limited preparation time.

2 Method

2.1 Participants

Six recruits in the marines' officer's training program POTOM (Praktische Opleiding tot Officier der Mariniers) participated in this study (age between 24 and 30 years, on average 27 years). Two participants were already employed at the Royal Netherlands Marine Corps (RNLMC) in an operational and supervisory function. This research was approved by the ethical committee of TNO.

2.2 Setting

This study was performed by four researchers during a live fire training of the troop hasty attack by the Royal Netherlands Marine Corps (RNLMC) in the hilly Sennybridge Training Area in Wales (UK) in May 2017. Within four days, six runs of hasty attacks were studied, four in the morning and two in the afternoon. Each participant held the position of troop commander during one run. A run consisted of one or more hasty attacks. Each run was studied by a pair of researchers.

2.3 Procedure

Upon arrival at Sennybridge, all participants were briefed on the purpose of the research. A day before the participant would perform the hasty attack run, the participant signed the informed consent form and was instructed to fill out the questionnaires regarding sleep duration and quality the next morning directly upon waking. Immediately before the hasty attack run took place, the participant also indicated his level of sleepiness. During the run, two researchers closely followed the troop commander. On average, a run lasted two hours. Directly after completing the run, the participant attended the after-action review held by the instructors, followed by filling out the cognitive and physical load questionnaires and an interview under the guidance of the two researchers. This interview took place in a private room to ensure the privacy of the participants. Finally, participants were thanked for their time and effort. On the same day as the hasty attack the POTOM lead instructor filled out a questionnaire rating the performance of the recruit.

2.4 Measures

In this study, a cognitive task analysis (CTA [5]) was used to examine the cognitive load of a troop commander during a hasty attack. The CTA consisted of three parts, namely observation of behavior during the hasty attack runs, an interview directly after the hasty attack run, and various questionnaires before and after the hasty attack run. As preparation, before the CTA took place, a hierarchical task analysis (HTA [6]) was conducted to obtain an understanding of the tasks and subtasks of a troop commander during a hasty attack. Doctrines, interviews with Subject Matter Experts and a field observation were used to gather this knowledge.

Observation. Two researchers observed whether and how the troop commander performed the tasks required during a hasty attack using an observation form based on the HTA. The researchers observed the troop commander from a short distance (~ 10 m) and were listening to either the troop radio net or the squadron radio net. After the hasty attack, the observers also attended the after-action review.

Interview. Directly after the after-action review, an interview was conducted with the participant using the Critical Decision Method (CDM [5, 7]) one of the most commonly used methods of implementing a CTA [8]. This method gives the interviewer the opportunity to return to a critical moment together with the interviewee and to identify why and based on what information, decisions were made [7]. The interview was structured using four phases. In the first phase a specific event was identified that caused a high level of cognitive workload for the participant. In the second phase, a timeline was constructed to obtain a clear overview of this event. In the third phase, the event was discussed in more detail to better understand how the situation and the cognitive load was experienced by the participant (e.g., what information was available, how did they make the decision). In the last phase, 'what if' questions were asked to invite the participant to speculate on how his decisions might have differed when having, for example, different technology available during the attack. Interviews lasted an hour. When there was time left after discussing one event, another event with a high cognitive load was identified and analyzed. The audio during the interviews was recorded.

Questionnaires. Participants filled out questionnaires regarding sleep since earlier research conducted with a previous POTOM class showed an accumulation of sleep debt [9]. Duration of sleep and subjective sleep quality of the previous night were measured with questions from the Pittsburgh Sleep Diary [10] and Karolinska Sleep Diary [11]. In addition, a question was asked about the average sleep duration during the whole training period in Sennybridge that started two weeks before the start of this study. At the beginning of the hasty attack run, participants filled out the Stanford Sleepiness Scale [12] to measure their level of sleepiness.

To gain insight into the experienced cognitive load during the hasty attack, participants completed the NASA TLX [13] and the Rating Scale Mental Effort (RSME [14]). To assess experienced physical load, participants filled outthe rate of perceived exertion (RPE) questionnaire [15].

Additionally, the POTOM lead instructor filled out a short questionnaire concerning his view on the recruit's performance. Five-point scales were used to indicate general functioning, stress resistance, learning ability and potential as troop commander. A one indicated 'very bad' and a five 'very good'. Furthermore, a grade was given between 1 and 10 to indicate how well the hasty attack run was executed by the recruit.

2.5 Data Analysis

Questionnaires. The results of the questionnaires filled out by the participants and the instructors were visualized in graphs and globally reported such that a participant's individual scores are no longer recognizable.

Observation and Interview. The transcripts of the interviews and notes from the observations were analyzed by the four involved researchers. Each researcher analyzed each hasty attack run (six in total) by answering the following seven questions:

- Which phase(s) in the hasty attack caused the highest cognitive workload?
- Which elements were contributing to this cognitive workload?
- How was the behavior of the participant (based on observation)?
- What were the operational consequences of the cognitive workload?
- What were the mitigation strategies applied by the participant?
- How did the participant expect that a more experienced troop commander would respond?
- What technologies could have supported the participant during the hasty attack?

Afterwards, these insights were discussed by the four researchers to create a common understanding of the experience of the participant (see Results section).

3 Results

3.1 Questionnaires

Sleep Duration and Sleep Quality. Table 1 shows the results of the questionnaires concerning sleep duration, sleep experience and the Stanford sleepiness. Results show that participants did have an acceptable sleep duration and experience.

Subjectively Experienced Load. Both physical and cognitive load were relatively high. The physical load (RPE) was on average 14.3 indicating 'relatively heavy' (*SD* = 1.5). The cognitive load calculated by averaging over the six NASA-TLX questions and participants was 13.7 (*SD* = 1.22) on a scale of 0 (none) to 20 (maximal). The average mental effort (RSME) score was 67.6 (*SD* = 19.3) on a scale of 0 to 150. Four participants indicated 'considerable effort', one participant 'great effort' and one participant 'some effort'.

Table 1. Results of sleep duration, sleep experience and Stanford Sleepiness scale

Question	Minimum	Maximum	Average
Sleep duration (night before)	4h25	7h15	6h24
Sleep duration (average across training)	5h	8h	6h30
How did you sleep	2	3	2.2
How refreshed after awakening	3	5	3.3
How quiet was your sleep	1	4	2.5
Slept throughout time allotted	3	4	3.2
Ease of waking up	2	3	2.5
Ease of falling asleep	1	5	2.5
Stanford Sleepiness scale	1	4	2

Performance Evaluation by Instructor. The participants were rated by the POTOM lead instructor on several scales, see Table 2. The highest scores on all scales belonged to the participants with previous military experience.

Table 2. Performance evaluation by instructor

Performance scale	Minimum	Maximum	Average
General functioning	2.75	5	3.5
Stress resistance	2.75	4	3.6
Learning ability	2	5	3.5
Potential as troop commander	2.75	5	3.5
Hasty attack score	5	9	6.7

3.2 Observation and Interview

For a better understanding of the results in this section, Table 3 gives an overview of the phases of a hasty attack. The next subsections will address the cognitive load during these different phases.

Table 3. Overview of tasks of the troop commander during a hasty attack

Phases of hasty attack	Tasks
1. Battle Preparation	The troop commander prepares his troop for the upcoming attack by determining marching orders and ensuring that his troop is prepared to react on enemy fire. He also informs the troop about the initial attack plan

(*continued*)

Table 3. (*continued*)

Phases of hasty attack	Tasks
2. Reaction on effective enemy fire (occurs every time that an enemy is encountered)	
2a. Initial reaction and assessing the situation	During this phase the troop moves forward to provoke enemy fire. When enemy fire is encountered the section under fire responds with a contact drill. The troop commander needs to move into a safe position where he obtains a good overview of the situation and gathers situational awareness (i.e. terrain, information about enemy, position own sections). He is in control of keeping the enemy under fire to start gaining back initiative in the fight and can therefore if needed assign extra sections. Besides that, he needs to brief the squadron commander on the situation
2b. Quick Estimate	The troop commander gains as much situational awareness as needed and comes up with a plan for the attack. He needs to decide on how to maneuver towards the enemy (frontal or left/right flank approach). Furthermore, he informs and asks approval of the squadron commander for his plan
3. Attack	
3a. Orders	The troop commander physically meets with the deputy troop commander and section commanders to brief them on the upcoming attack. Afterwards, the section commanders will brief their sections
3b. Approach	A section of the troop that is not under fire will make a covert flanking or frontal movement towards the enemy. In the meanwhile, another section of the troop, keeps the enemy under fire, such that the maneuvering section can safely move forward. When the maneuver section arrives at a certain predefined location the fire needs to be stopped, to prevent friendly fire. This process is coordinated by the troop commander. In general, the troop commander joins the maneuver section, however, at a safe distance behind the section
3c. Assault	The maneuver section runs towards the position of the enemy and uses firepower to subdue them. This can be coordinated either by the troop commander or the section commander
4. Reorganization	When the enemy is defeated the troop needs to quickly set up adequate security in the area. The troop commander coordinates this process and needs to determine whether the troop can continue with the fight, needs replenishments of goods, or needs to be relieved by another troop. Additionally, he needs to give a situational report to the squadron commander

Cognitive Demanding Phases During the Hasty Attack. Table 4 shows the phases during the hasty attack that were highlighted as most cognitively demanding by the participants during the interviews. The number of hasty attacks a participant performed during a run varied per participant. When multiple enemies were encountered, the later encounters seemed to be more cognitively demanding than the first. During the later encounters, the troop was often more scattered over the terrain due to previous attacks and the evacuation of casualties.

Table 4. Most cognitive demanding phases during the runs as indicated by the participants.

Participant	Number of attacks	2a Initial reaction	2b Quick estimate	3a Orders	3b Approach	3c Assault	4 Reorg
1	4	4th contact			3rd contact		
2	2		1st contact				
3	1		1st contact		1st contact		
4	2	2nd contact			1st contact		2nd contact
5	1		1st contact				1st contact
6	4		4th contact				4th contact

Four phases were indicated as most demanding by the participants:

- 2a Initial reaction: This phase was indicated as most demanding in the situation of a sudden new enemy contact, while the reorganization after previous enemy contact (s) was not yet completed. During these high demanding initial reactions, the troop was relatively dispersed over the terrain due to multiple attacks or casualties. Maintaining command and control was reported as more difficult.
- 2b Quick estimate: During this phase the highest cognitive demand was during planning of the hasty attack. This demand was high for several reasons; difficulties to decide how to approach the enemy, e.g., via the left or the right flank, the multitude of events occurring simultaneously, e.g. casualties, or a lack of basic skills needed to communicate the plan according to protocol to the squadron commander.
- 3b Approach: Two different high demanding situations were reported during the approach phase. First, during two hasty attacks, the troop commander misinterpreted the terrain in terms of opportunities to hide for the enemy, resulting in a situation where their own troops were visible for the enemy. Therefore, a new plan was made and communicated. During another hasty attack, the troop did not

approach the enemy from the correct angle, causing a potentially dangerous situation increasing the risk of friendly fire.

- 4 Reorganization: Participants had difficulties with setting up a good strategic position in the terrain (i.e., a position with a clear 360° overview of the surrounding area and hidden from the enemy). In addition, it was difficult to keep the priorities in the right order, e.g., prioritizing safety of the troop above dealing with casualties.

Cognitive Workload Factors. The different elements contributing to the cognitive load were categorized in different cognitive load factors. Table 5 gives a description of these factors and in how many events these factors contributed to the cognitive load. The factors are ordered by the number of events.

Table 5. Overview of factors that increased the cognitive workload during the events mentioned as cognitive demanding.

Cognitive load factor	Description	Number of events
Command and control (C2)	Keeping the squadron commander up to date of the troop's status and controlling the hasty attack executed by the troop. This includes creating an overview of the locations and status of all the sections, controlling the sections with the right priorities (safety first), and ensuring that the sections understood and executed the actions ordered	9
Threat	Unexpected dangerous threats such as an armed vehicle that was approaching or being in a dangerous position (without protection) in the terrain	5
Casualties	When sections were in field of view of the enemy or a mistake was made by one of the sections, casualties were assigned by the instructor. The deputy troop commander was responsible for dealing with these casualties and transporting the wounded away from the battlefield, while the troop commander had to maintain safety and coordinate the attack with less men available. Also, the screaming of the wounded, could distract the troop commander	5
Terrain	Various factors made interpreting the terrain difficult. First, the terrain was very open making maneuvering difficult. Secondly, the terrain was full of hills causing difficulties with estimating distances, finding spots with line of sight and cover, and planning the attack as the troop commanders did not have much experience yet with hilly terrain	5
Education/instructors	The pressure to perform well to stay in the officers' training program and pressure from the instructors during the hasty attack (asking questions caused doubts about decisions)	4

(*continued*)

Table 5. (*continued*)

Cognitive load factor	Description	Number of events
Information	Large amount of information needs to be processed simultaneously including irrelevant information	2
Time pressure	Keeping momentum during the hasty attack	1
Planning	Coming up with the plan for the hasty attack	1
Sound	Hearing much sound from the environment such as people shouting and shooting and sound coming in over multiple radios	1
Military basic skills	Possessing basic military skills such as determining the position of the enemy on a grid	1

Figure 1 gives an overview of the different cognitive load factors that occur within the different phases of the hasty attack. The most cognitive load factors were experienced during the quick estimate phase; nine out of ten workload factors were contributing to the workload in this phase. A factor contributing to all phases were the casualties.

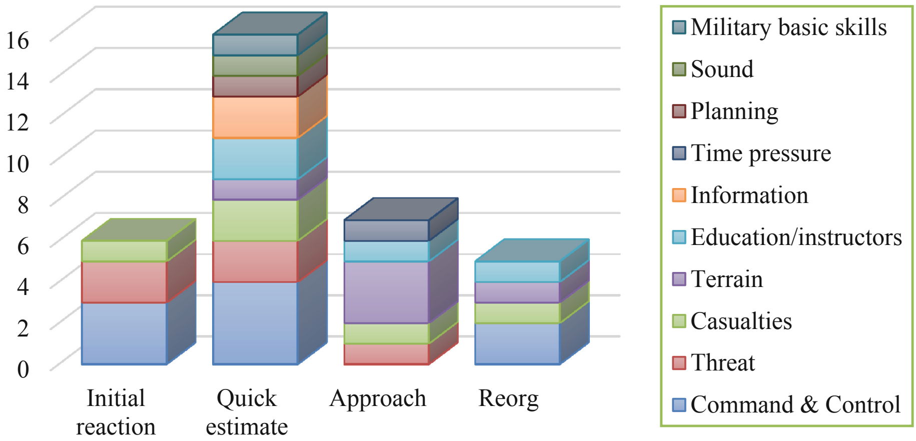


Fig. 1. Occurrences of cognitive load factors for each phase in the hasty attack

Mistakes During the Hasty Attack and Consequences. Below an overview of mistakes of the recruits that we identified from the observations, interviews, and after-action reviews ordered by cognitive load factor.

Command and Control. Mistakes resulting from a lack of command and control typically result in reduced safety, reduced sustainability and loss of momentum of the attack. Typically mistakes were: a lack of awareness of the position of all sections, positioning themselves near the front-line when confronted with an armored threat, forgot to allocate the weapon systems during the reorganization and forgot to clearly instruct the attack plan to a section.

Terrain. Amistake that was made by multiple recruits was choosing the wrong route for the flanking maneuver due to inaccurate estimation of terrain characteristics. In one occasion, this reduced sustainability and momentum as the recruit should create a new plan. In another case, the recruit continued the operation through open terrain and therefore increasing the risk being detected by the enemy.

Threat. In several runs, recruits forgot their own cover in the field (e.g., lifting their heads too high or being alone in the field without a section for protection). This is very risky as it is important for the mission and troop that the troop commander remains safe.

Basic Skills. One recruit gave a wrong coordinate of the enemy position. This mistake was corrected by the instructor, but otherwise it could have increased the risk of friendly fire and not being able to defeat the enemy during the attack.

Supporting Technologies. During the interviews, the recruits were asked which technologies would support the execution of the hasty attack. Several technologies were mentioned. To increase situational awareness, a tool that tracks your own position, the position of the other sections, or the enemy's position was mentioned. This information can be used during the quick estimate phase. Also, eyes in the sky (i.e., drones) could help them to find the enemy location or to explore terrain characteristics (e.g., finding locations to cover for the enemy), while the troop commander can stay safely in the back of the terrain. Augmented reality was mentioned as technology that allow them to see through the eyes of the fire section (this section is often closer to the enemy) resulting in an increase in situational awareness. In addition, one recruit suggested that inclinations of the terrain should be visible on a map. Altitude lines on a map can be difficult to interpret and can make it harder to estimate distances. Another recruit suggested a digital foldable map allowing them to zoom in and out or see the terrain from another perspective. Smaller radios were also suggested, as the radio currently carried by the troop commander is large and limits their mobility. Overall findings showed that the most frequently mentioned technologies were those that increase situational awareness.

4 Discussion

In the current study, the CTA methodology was applied to a troop hasty attack aiming to better understand the cognitive load of a troop commander during such an attack. Results showed a relatively high level of cognitive load and different phases of the attack contributing to this load. The phases that contributed most to the experienced load were the initial reaction, quick estimate, approach and reorganization. For each phase, the cognitive load factors contributing to the cognitive load varied. The number of cognitive load factors was the largest during the quick estimate, where a variety of cognitive tasks were performed by the troop commander. During the quick estimate, four crucial cognitive tasks occurred simultaneously, namely acquiring situational awareness, deciding the best route for approaching the enemy, and communicating and planning the attack. The results of the interview showed that during this phase also several mistakes were made that could have been the result of the high cognitive load experienced by the recruits. Typical mistakes were: choosing the wrong route, unsafe positioning during information collection (i.e., visible for the enemy or too close to contact), unclear orders during the communication of the plan and inaccurate awareness of the position of the sections in the field. In the current study, the assault phase was not mentioned as a high demanding phase. However, during an attack in real life, the experienced cognitive load will probably be larger, due to higher levels of experienced stress that is difficult or even impossible to simulate during a training.

The factor most frequently mentioned to contribute to the cognitive load was Command and Control (C2), which is also an important task for a troop commander during a hasty attack. Two other factors, casualties and threat, were also mentioned frequently. These factors were demanding for the recruits since both factors happen suddenly and most of the time the original plan needed to be redefined. For example, in a threatening situation, the threat should be eliminated as quickly as possible while taking into account the safety of the troop. Finally, also the terrain was mentioned as a demanding factor. Perceiving and understanding the characteristics of the terrain is an important prerequisite to ensure a safe and efficient elimination of the threat. Many recruits had difficulties with the correct interpretation of the terrain, finding the optimal route to attack the enemy and to position themselves having a clear overview of the battle space without being seen by the enemy.

The current study also has some limitations. The first limitation is that during the training program special attention was paid to C2 and the recruits knew they were credited based on their performance, that could have overestimated the contribution of C2 on cognitive load. In addition, the presence of instructors and the fact that the performance of the recruits was judged might have changed their behavior and decisions. Another factor is the number of updates asked by the squadron commander. This was more than during a regular training or a real life hasty attack. Additionally, the recruits that were assigned to be one of the section leaders also tried to show the best of themselves and gave more information on the radio than in a more realistic scenario. This might have increased cognitive load, since the troop commander had more information available to process.

5 Mixed Reality Platform

The different insights gained in the current study are used for the development of a mixed reality platform; a platform that combines new technological innovations with a virtual reality environment in which users will interact with their physical equipment while walking on a treadmill that is coupled to a virtual world displayed on a large screen to increase immersion (see Fig. 2). Such a platform offers the opportunity to investigate the effects of new technological innovations on cognitive performance under conditions mimicking real-life situations, while controlling for potential confounding variables. In this section, the important insights are discussed that are prerequisites for the development of such a platform.

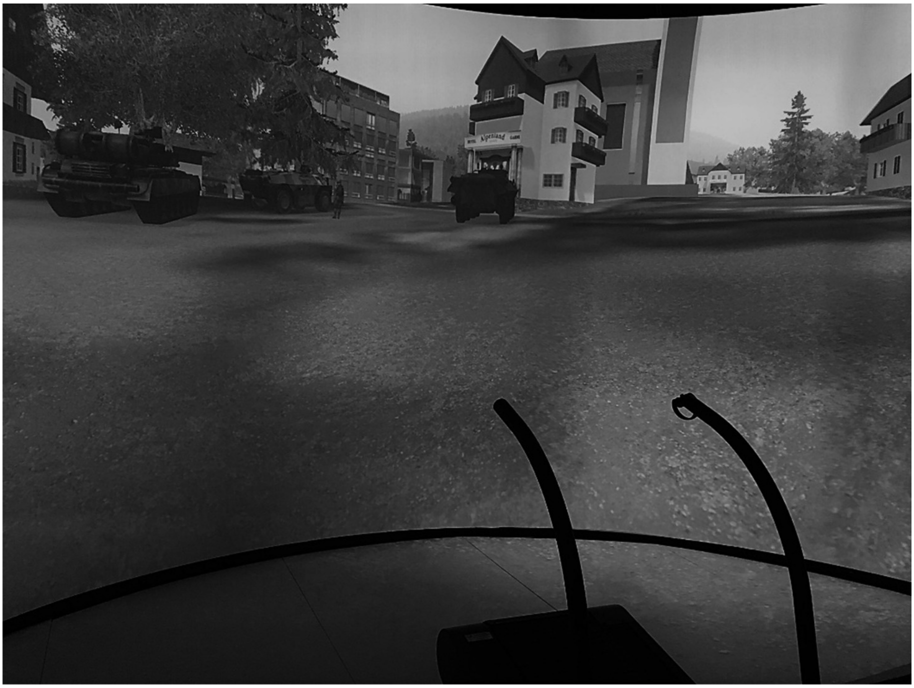


Fig. 2. Example of the mixed reality platform

5.1 Cognitive Load

From the four most cognitive demanding phases the quick estimate and approach are the most interesting to implement in the mixed reality platform. During the quick estimate the troop commander was applying all the steps in the situational awareness (SA) cycle as defined by [16]; perception, comprehension and projection. According to [16] and in line with our observations, the SA the commander has of the battle field (e.g., where is the threat, where are my sections, characteristics of the terrain) determines the decisions the troop commander is making (e.g., what is the best route to follow, how to position my sections, what weapons systems do I need). During the approach phase the troop commander was often confronted with unexpected events increasing cognitive load (e.g., the terrain was more open with less elements to hide, another enemy appeared, casualties occurred). Therefore, the virtual environment and the scenario should be realistic. For example, the terrain must provide possibilities to hide and overlook the battle field. In addition, for a realistic situation, relevant stressors that impact cognitive performance such as environmental noise, time pressure, uncertainty, fatigue, and unexpected events must be implemented in the scenario.

Both during the quick estimate and approach, the troop commander performed many different tasks and frequently switched between them (e.g., inform higher command, dealing with casualties, estimate the (new) enemy position, preparing the attack, change the initial plan based on new information). Therefore, the scenario in the mixed reality platform should include multiple tasks that are performed simultaneously.

Communication with the section and squadron commander is important to gain situational awareness and control the troop. When an order is communicated to the sections, the execution of that order should be monitored carefully to ensure the plan is correctly executed, which increases cognitive load. In addition, when the plan is not well understood it will increase the likelihood that the troop commander must intervene, which costs additional cognitive resources and time. Furthermore, environmental noise, such as gun shots, will increase the difficulty to understand messages, resulting in an increase of cognitive load. Therefore, within the mixed reality platform the commander should be able to communicate with at least two section commanders (e.g., one section that is under fire and another section that approaches the enemy). Hence, communication devices will be coupled to allow the troop commander to communicate with other sections and the squadron commander.

5.2 Technologies

Numerous technologies are currently on the market that claim to increase operational performance. The CTA showed that technologies that support the commander to build up SA have large potential. Examples are systems with blue force tracking or systems with cameras to inspect the terrain. Integrating these into the mixed reality platform allows us to study how and when information can be presented most effectively, which information displays are most effective (e.g., hand held displays, augmented reality) or what modalities can be used in this context (e.g., haptic, visual, auditive).

5.3 Physical Exertion

The troop commander experienced a high level of physical load before the quick estimate phase and during the approach phase, since the troop commander followed the section that approached the enemy. This stresses the importance to study cognitive performance while being physically active or induce physical fatigue before cognitive demanding tasks. Therefore, a treadmill will be coupled to navigate within the virtual environment.

6 Conclusion

The results of the current study successfully contributed to the current development of a mixed reality platform. In the future, this platform allows to examine the effects of new technological innovations on cognitive performance under conditions mimicking real-life situations, while controlling for potential confounding variables. In the future, it would be interesting to perform a cognitive task analysis with more experienced military and include a larger set of military tasks. This might result in additional requirements that can be used to optimize the mixed reality platform. In addition, the mixed reality platform will be developed and tested involving experienced military users to ensure operational relevancy.

Acknowledgement. This work was supported financially by the Dutch Ministry of Defense (V1605 program).

References

1. O'Donnell, R.D.O.: The Army Cognitive Readiness Assessment (ACRA) System – A Neuroergonomic Approach (2011)
2. Rice, V.J., et al.: Automated Neuropsychological Assessment Metrics (ANAM) Traumatic Brain Injury (TBI): Human Factors Assessment (2011)
3. Tack, D.W., Angel, H.: Cognitive task analyses of information requirements in dismounted infantry operations (2005)
4. Wickens, C.D., Lee, J.D., Liu, Y., Becker, S.E.G.: An Introduction to Human Factors Engineering, 2nd edn. Addison-Wesley and Longman, New York (2011)
5. Crandall, B., Klein, G., Hoffman, R.R.: Working Minds: Cognitive Task Analysis. MIT Press, Cambridge (2006)
6. Stanton, N.A.: Hierarchical task analysis: developments, applications, and extensions. Appl. Ergon. **37**, 55–79 (2006). https://doi.org/10.1016/j.apergo.2005.06.003
7. Klein, G.A., Calderwood, R., Macgregor, D.: Critical decision method for eliciting knowledge. IEEE Trans. Syst. Man Cybern. **19**, 462–472 (1989). https://doi.org/10.1109/21.31053
8. Plant, K.L., Stanton, N.A.: What is on your mind? Using the perceptual cycle model and critical decision method to understand the decision-making process in the cockpit. Ergonomics **56**, 1232–1250 (2013). https://doi.org/10.1080/00140139.2013.809480
9. Simons, M., Valk, P.J.L., Vrijkotte, S., Veenstra, B.J.: Performance and Health monitoring during a Marines Training Course R11469 (2013)
10. Monk, T.H., et al.: The Pittsburgh sleep diary. J. Sleep Res. **3**, 111–120 (1994)
11. Åkerstedt, T., Hume, K., Minors, D., Waterhouse, J.: The meaning of good sleep: a longitudinal study of polysomnography and subjective sleep quality. J. Sleep Res. **3**, 152–158 (1994). https://doi.org/10.1111/j.1365-2869.1994.tb00122.x
12. Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., Dement, W.C.: Quantification of sleepiness: a new approach. Psychophysiology **10**, 431–436 (1973). https://doi.org/10.1111/j.1469-8986.1973.tb00801.x
13. Hart, S.G., Staveland, L.S.: Development of the NASA-TLX: results of empirical and theoretical research. In: Human Mental Workload, pp. 139–183. Elsevier (1988)
14. Zijlstra, F.R.H.: Efficiency in work behavior. A design approach for modern tools. Delft University of Technology (1993)
15. Borg, G.A.V.: Psychophysical bases of perceived exertion. Med. Sci. Sports Exerc. **14**, 377–381 (1982). https://doi.org/10.1249/00005768-198205000-00012
16. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Hum. Factors: J. Hum. Factors Ergon. Soc. **37**, 32–64 (1995). https://doi.org/10.1518/001872095779049543

Impacts of Automation Reliability and Failure Modes on Operators' Performance in Security Screening

Zijian Yin(✉), Pei-Luen Patrick Rau, and Zhizhong Li

Department of Industrial Engineering, Tsinghua University,
Beijing 100084, People's Republic of China
yinzjl8@mails.tsinghua.edu.cn,
rpl@mail.tsinghua.edu.cn, zzli@tsinghua.edu.cn

Abstract. Nowadays more and more automated systems are used to help human operators to finish their tasks. In security screening of mass transit systems, there have been many studies focused on automated detection, but few can reach completely reliable. It is quite often that human operators work together with not-completely reliable automated systems. This study investigates the influence of automation reliability and failure modes on operators' performance in security screening of mass transit systems. Results show that higher reliability and failure modes of "false alarms" can improve security screening performance, while the latter has some negative impacts on efficiency.

Keywords: Security screening · Automated detection system · Reliability · Failure modes

1 Introduction

With the rapid growth of population in the world, mass transit systems are used broadly, such as railways and subways. In Beijing, for example, buses and subways become more and more preferred by many citizens, since using mass transit systems are not only cheap but also more efficient in rush hours. However, as the number of people choosing mass transit systems increases, such systems have become the targets of terrorist attacks around the world, because attacks in mass transit systems will arise much attention and cause public panic (Fiondella et al. 2013). Thus, a reliable security screening system is essential to guarantee the safety of public transit.

In current security screening systems, detecting technology is mainly based on scanning and imaging of the baggage, like X-ray. Via the scanning technology like X-ray, the image of the baggage will be presented in computers as well as items in the baggage, and these images are checked by operators who are responsible for monitoring each bag whether there exist forbidden items in it. In general, forbidden items mainly include cutleries, flammable things, explosives, guns, etc., depending on related laws and regulations of each country. Thus, operators' workload is not low, as both of the security and the efficiency of mass transit systems must be ensured. It is also a complex work for operators to identify so many kinds of forbidden things accurately. If

The original version of this chapter was revised: Reference 6 has been corrected. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-22507-0_34

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 137–149, 2019.
https://doi.org/10.1007/978-3-030-22507-0_11

one operator is not sure about whether a forbidden item exists or not in a specific bag, he/she needs to suspend the conveyor, and open the bag physically to check again. In a word, the responsibility of the operators in security screening is to monitor the bags taken with passengers and find out forbidden items from them, which is not a simple task.

However, the passenger flow in mass transit systems is heavy, especially in rush hours and holidays. If every bag is checked thoroughly, the service rate will decrease sharply, which will cause a large number of passengers delayed in the system. This result has been proved in Fiondella et al.'s study (2013). Therefore, too rigorous security check procedures are not suggested, since this will influence the normal function of mass transit systems.

Applying automated detection systems in security screening checkpoints is an available way to help operators improve their performance in screening work. Automated detection systems can identify potentially dangerous items in an image, based on machine learning on given training sets. Current automated detection systems can detect the most dangerous items, but not everything. Moreover, it is difficult for automated detection systems to detect those type of items which have not been trained before. When working with this not-completely reliable system, the operator's performance may be affected. This study aims to investigate the influence of automation reliability and failure modes on operators' performance in security screening of mass transit systems.

The remainder of this article is organized as follows. Section 2 reviews some current research about machine learning, human-automation system cooperation, and automation reliability. Section 3 describes our methodology used to explore the impacts of automation reliability and failure modes on operators' performance. Section 4 reports the result of data analysis. Section 5 discusses the result and gives some implications for the design of automation detection algorithm. Section 6 reaches a conclusion of this study.

2 Literature Review

2.1 Reliability of Automated Detection

There have been large amounts of studies focused on the development of automated detection algorithms. For example, Mery et al. (2016) proposed an algorithm named Adaptive Sparse Representation (XASR+), which consisted of two stages: learning and testing. The algorithm was tested for the detection of 4 different objects, and all recognition rates were more than 95% (Mery et al. 2016). Kevin (2018) developed convolutional object detection algorithms trained on annotated x-ray images and tested the algorithm on images of luggage. Results indicated that this algorithm could detect selected forbidden items with high accuracy (95.5% for firearms and 94.0% for sharps) and small impact on false alarm rates (1% for firearms and 3% for sharps) (Kevin 2018).

From these results reported, we can see that current automated diagnostic algorithms have high accuracy but still cannot reach 100%. Therefore, exploring the situation where human works with not-completely reliable automation systems is valuable.

2.2 Modes of Detection Failures

Based on the signal detection theory, Dixon and Wickens (2006) pointed out that, in application, the results of imperfect detection system could be divided into two categories: “misses” and “false alarms”, while different values of “β” (the threshold of the detection algorithm) would contribute to different failure modes. When the “β” value was set higher, fewer alarms would generate, and the ratio of failure mode “misses” would increase (Dixon and Wickens 2006). On the other hand, the ratio of failure mode “false alarms” would increase when the “β” value was set lower (Dixon and Wickens 2006). Too many false alarms would destroy human trust in automation, while too many misses would require more human effort and then turn down the benefit of automation. It is reasonable to hypothesize that failure modes of automated detection would influence operator performance in security screening.

2.3 Criteria to Evaluate the Screening System

Fiondella et al. (2013) has proposed a method to assess the performance of mass transit systems. In Lance Fiondella’s assessing model, two important parameters were considered: the security of the mass transit system and delays incurred on the traveling flow (Fiondella et al. 2013). “Security” referred to the ability of the screening system to detect forbidden items in baggage. The designed function of screening systems was to identify all possible illegal and forbidden items to ensure the safety of the system and people. Thus “security” was the primary attribute when evaluating a screening system. Meanwhile, the negative impact caused by screening is the decreasing of efficiency of the mass transit system. If the screening procedure causes substantial delay, then this detection system needs proving.

3 Method

3.1 Independent Variables

In this study, two independent variables were considered, i.e. reliability and failure modes of automated detection system.

Reliability of Automated Detection Systems. According to Sect. 2.1, it is common that operators have to work with this not-completely reliable system. The reliability of the automated system could have some influence on operators’ performance (Dixon and Wickens 2006; Singh et al. 2009; Rice and McCarley 2011; Hillesheim and Rusnock 2017; Rovira et al. 2007). So this independent variable aimed to evaluate the influence of the reliability on operators’ performance, considering the context of security screening in mass transit systems. Here, the value of the reliability was calculated in Eq. (1). And there were two levels of this independent variable: 70% and 90%.

$$\text{Re}liability = \frac{\#\,of\ correct\,detection\ by\,automation\,system}{\#\,of\ detection\ by\ automation\ system} \times 100\% \tag{1}$$

Failure Modes. Failure modes represented what kind of errors could occur when the automated detection system functioned. Based on Dixon and Wickens's study (2006), we designed two levels of this independent variable: misses and false alarms. If in one treatment the failure mode was "misses", then all possible errors would be misses, and no false alarm would occur.

3.2 Dependent Variables

In this study, six dependent variables were considered.

CR
Correct detection rate of the participant, defined as Eq. (2).

$$CR = \frac{\#\,of\ correct\ detection}{Total\,\#\,of\ cases} \tag{2}$$

MR
Misses rate of the participant, defined as Eq. (3).

$$MR = \frac{\#\,of\ misses}{\#\,of\ target - present\ cases} \tag{3}$$

FR
False alarms rate of the participant, defined as Eq. (4).

$$FR = \frac{\#\,of\,false\ alarms}{\#\,of\ target - absent\ cases} \tag{4}$$

PRT
Mean reaction time of the participant in target-present cases (ms). According to Chun's study (1996), the human searching process under target-present cases and target-absent cases is not the same. So the reaction time under two cases is measured and analyzed separately.

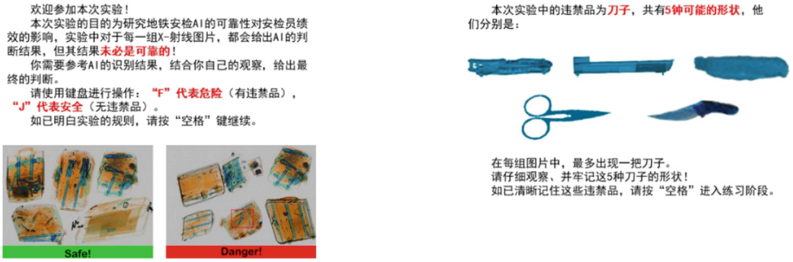
ART
Mean reaction time of the participant in target-absent cases (ms).

3.3 Participants

There were 12 participants invited to this experiment. All of them were undergraduate students from Tsinghua University (10 males & 2 females, mean age = 23.2). The participants were all right-handed, and with normal or corrected-to-normal vision. All participants had no occupational experience in security screening check, so there was no considerable difference among their skills in security screening.

3.4 Experiment Design

The experiment was constructed as a 2 * 2 full factorial design, and within-participant design for the two independent variables. In the beginning, the participant was shown with a guidance interface, which informed the purpose and the task to the participant, shown in Fig. 1. When they were reading these instructions, we would emphasize for participants that in this experiment, the forbidden items only included these five types of cutleries; if any of them appeared in the image, then this image was dangerous. In each image, there was four luggage, and any type of cutlery could appear in any position in any of the four luggage.



欢迎参加本次实验！

本次实验的目的为研究地铁安检AI的可靠性对安检员绩效的影响，实验中对于每一组X-射线图片，都会给出AI的判断结果，但其结果未必是可靠的！

你需要参考AI的识别结果，结合你自己的观察，给出最终的判断。

请使用键盘进行操作：“F”代表危险（有违禁品），“J”代表安全（无违禁品）。

如已明白实验的规则，请按“空格”键继续。

本次实验中的违禁品为刀子，共有5钟可能的形状，他们分别是：

在每组图片中，最多出现一把刀子。
请仔细观察、并牢记这5种刀子的形状！
如已清晰记住这些违禁品，请按“空格”进入练习阶段。

Fig. 1. Instruction

After the participant had known the main content of this experiment, they were guided into a training procedure with nine trials, shown in Fig. 2. As shown in Figs. 3 and 4, the automation system would give a judgment to this image. If the system thought there existed dangerous items in one image, a mark “Dangerous!” was given below the image with a red alarm color. Otherwise, a mark “Safe!” was given below the image with a green color. In the training procedure, all judgments made by the automation system were right.

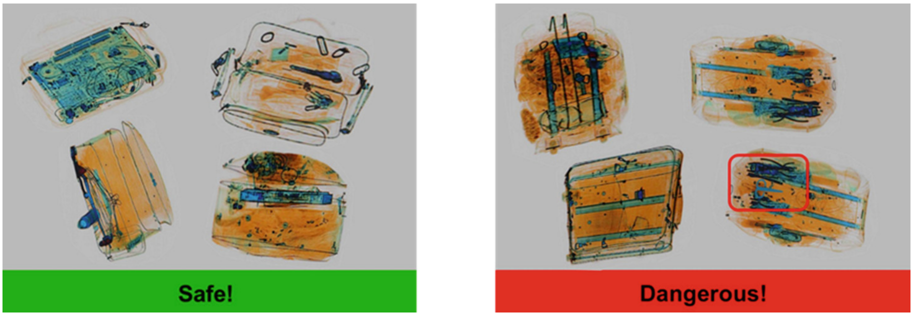


Fig. 2. Safe and dangerous images in the training (Color figure online)

After the training stage, the participant would start the formal experiment. The image shown in the formal experiment was similar to that in the training stage, except for that the reliability of the automation system was not 100%. In the formal experiment, there were totally four treatments (2 * 2 levels), and in each treatment 80 trials were completed by participants. Between every treatment's beginning, the participant was told to have a rest, until they felt no fatigue. In each trial, one image was presented to the participant, and he/she was required to judge whether this image was safe or dangerous. There was no time limit, and the participant needed to press "J" or "F" in the keyboard to give his/her judgment. Once a "J" or "F" was pressed, no reaction feedback about whether this decision was correct or not was given (see Carryover Effects in Sect. 3.5), and after an interval of 0.5 s the next trial (image) was given. Because no feedback was given, in the participant's view there should be no difference among these four treatments. Hence, just after each treatment, the participant was asked to fill up NASA-TLX to validate their similar experience among four treatments. The whole experiment process lasted about 30 min on average, and each participant finishing the experiment would receive ¥40 as remuneration for his/her attending.

3.5 Bias Controlling

Practice Effects

The two independent variables followed a within-subjects design, and every participant was required to finish all treatments. So practice effects could not be ignored as participants tended to be more familiar and skilled about this task. To reduce the practice effects in the within-subjects design, a Latin-square design was used to balance the order of treatments.

Fatigue Effects

There were totally 4 treatments and 80 trials in each treatment. So in the formal experiment, there were 320 images for participants to judge. To reduce the effect caused by fatigue, the participants were given enough rest between each turn, until they felt OK.

Carryover Effects

As participants might feel some characters of the task, they might change their strategies to obtain better performance. Besides, some studies had also discussed the impacts of "trust" and "compliance" (Rice and Mccarley 2011; Fiondella et al. 2013), and these could also lead to some biases in results. Thus, during the formal experiment, no feedback of the correctness of participants' answers was given to them.

Handedness Effects

In the experiment, participants needed to press "J" and "F" in the keyboard to give their judgment, and this could be influenced by their handedness. Thus, each participant was allocated into left-hand-safe or right-hand-safe groups at random.

Other Effects

The appearing position of forbidden items in images was randomized. The order of target-present cases and target-absent cases was randomized.

4 Results

A series of repeated-measures ANOVA was conducted to test whether significant differences existed. Shapiro-Wilk test was conducted for all values of dependent variables, and results showed that not all dependent variables' normality can be guaranteed. However, considering that the group sizes for conducting ANOVA are equal, the F-statistics can be robust to violations of normality (Donaldson 1968; Lunney 1970).

4.1 Correct Detection Rate

The mean correct detection rate under each condition is shown in Fig. 3. And the result of ANOVA is shown in Table 1. From the result of ANOVA, main effects of reliability (F(1, 11) = 10.40, $\eta^2 = 0.10$, p = 0.008) and failure modes (F(1, 11) = 5.81, $\eta^2 = 0.05$, p = 0.035) were significant and no significant interaction effects were found.

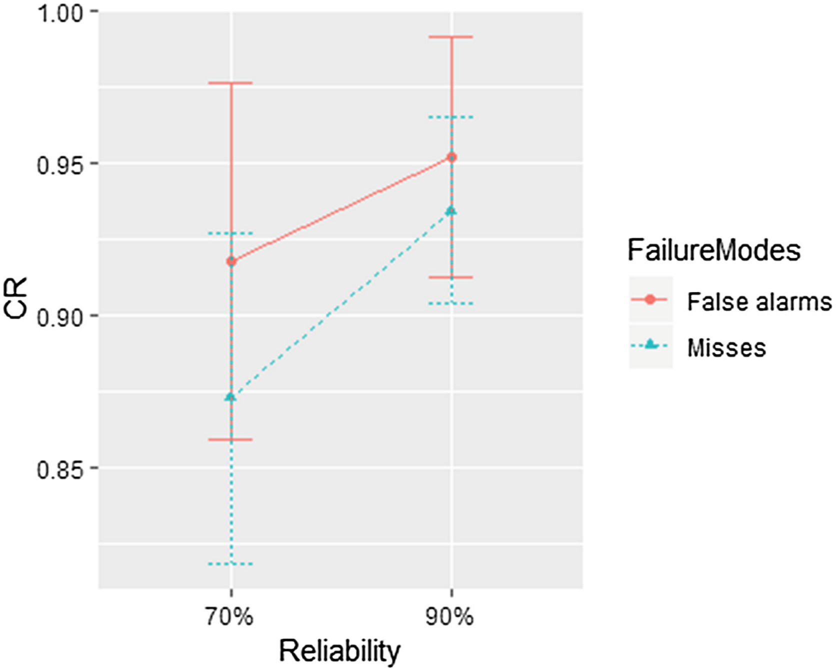


Fig. 3. Correct detection rate

Table 1. ANOVA for correct detection rate

Effect	DFn	DFd	SSn	SSd	F	p-value	η^2
(Intercept)	1	11	40.56	0.17	2669.01	<0.001*	0.99
reliability	1	11	0.03	0.03	10.39	0.008*	0.10
failure mode	1	11	0.01	0.02	5.81	0.035*	0.05
reliability:mode	1	11	<0.01	0.02	1.09	0.319	0.01

*: p < 0.05

4.2 Misses Rate

The mean misses rate under each condition is shown in Fig. 4. And the result of ANOVA is shown in Table 2. From the result of ANOVA, main effects of reliability (F (1, 11) = 8.37, η^2 = 0.09, p = 0.015) and failure modes (F(1, 11) = 17.62, η^2 = 0.15, p = 0.001) were significant, as well as the interaction effect (F(1, 11) = 5.06, η^2 = 0.03, p = 0.046).

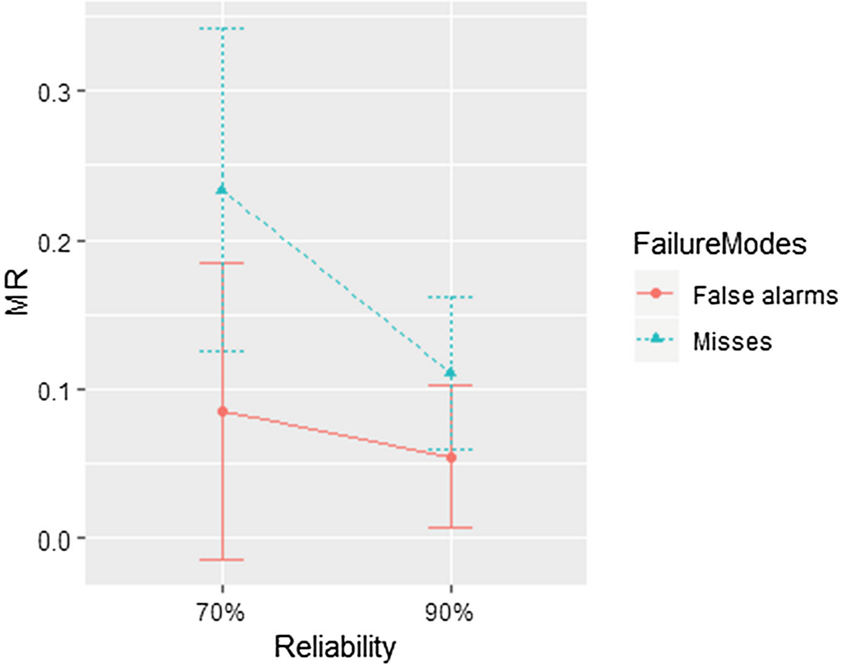


Fig. 4. Misses rate

Table 2. ANOVA for misses rate

Effect	DFn	DFd	SSn	SSd	F	p-value	η^2
(Intercept)	1	11	0.70	0.50	15.54	0.002*	0.49
reliability	1	11	0.07	0.09	8.37	0.015*	0.09
failure mode	1	11	0.13	0.08	17.62	0.001*	0.15
reliability:mode	1	11	0.03	0.06	5.06	0.046*	0.03

*: $p < 0.05$

4.3 False Alarms Rate

The mean false alarms rate under each condition is shown in Fig. 5. And the result of ANOVA is shown in Table 3. From the result of ANOVA, no significant effect was found, while the main impact of failure mode was marginal significant (F(1, 11) = 4.60, η^2 = 0.11, p = 0.055).

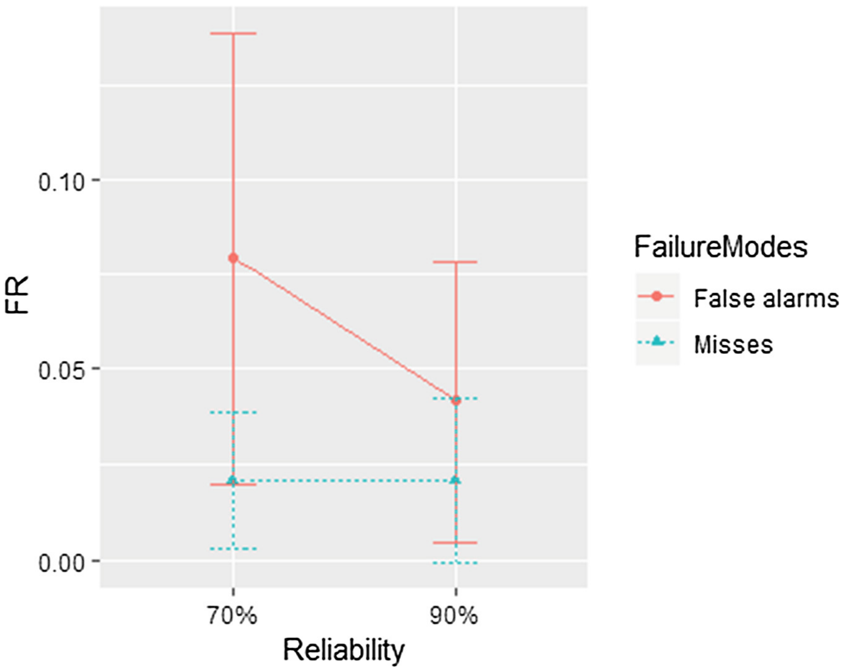


Fig. 5. False alarms rate

Table 3. ANOVA for false alarms rate

Effect	DFn	DFd	SSn	SSd	F	p-value	η^2
(Intercept)	1	11	0.08	0.05	18.29	0.001*	0.34
reliability	1	11	<0.01	0.04	1.19	0.298	0.03
failure mode	1	11	0.02	0.04	4.60	0.055^{marginal}	0.11
reliability:mode	1	11	<0.01	0.02	2.11	0.175	0.03

*: $p < 0.05$

4.4 Mean RT in Target-Present Cases

The mean reaction time in target-present cases under each condition is shown in Fig. 6. And the result of ANOVA is shown in Table 4. From the result of ANOVA, the main effect of reliability was significant (F(1, 11) = 7.69, $\eta^2 = 0.08$, p = 0.018).

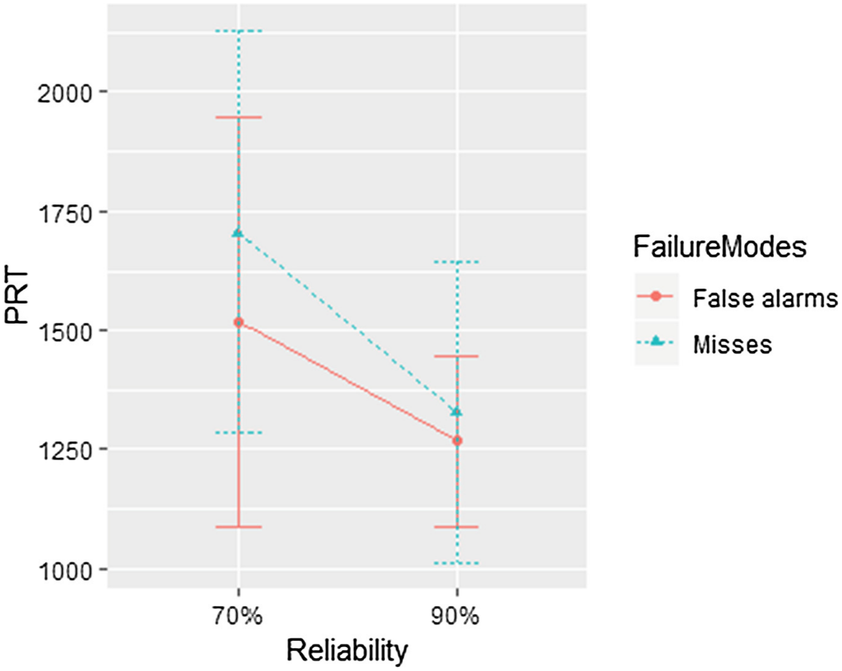


Fig. 6. Mean RT in target-present cases

Table 4. ANOVA for mean RT in target-present cases

Effect	DFn	DFd	SSn	SSd	F	p-value	η^2
(Intercept)	1	11	1.02×10^8	4.14×10^6	269.86	<0.001*	0.88
reliability	1	11	1.18×10^6	1.69×10^6	7.69	0.018*	0.08
failure mode	1	11	1.92×10^5	3.63×10^6	0.58	0.461	0.01
reliability:mode	1	11	4.86×10^4	4.06×10^6	0.13	0.724	<0.01

*: $p < 0.05$

4.5 Mean RT in Target-Absent Cases

The mean reaction time in target-absent cases under each condition is shown in Fig. 7. From the result of ANOVA, no significant effect was found.

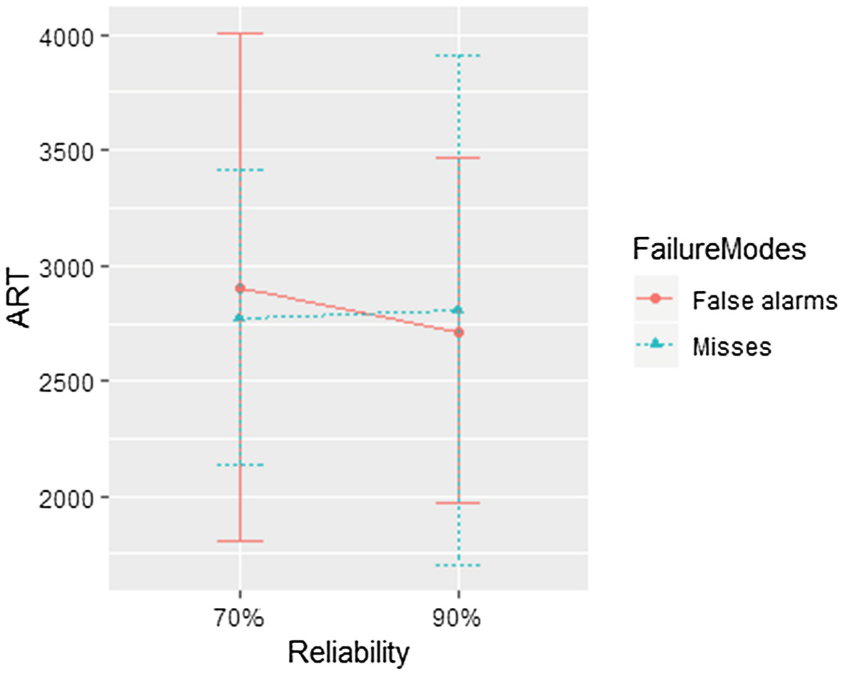


Fig. 7. Mean RT in target-absent cases

4.6 NASA-TLX Scores

The NASA-TLX scores under each condition is shown in Fig. 8. From the result of ANOVA, no significant effect was found. This result indicates that the no-feedback design functioned and participants felt the same workload among the four treatments.

5 Discussion

5.1 Security Related Items: Correct Detection and Misses

As mentioned in Sect. 2.3, there are two main criteria to evaluate a screening system: security and efficiency. Among the measured dependent variables, the correct detection rate and misses rate have more relationships with security, as misses will directly cause those illegal items brought into the system, which is very dangerous. From the results of ANOVA, for misses rate, all main effects were significant. This indicates when the reliability of systems increases, the misses rate decreases. Systems with higher reliability can provide sufficient help to operators; images including forbidden items are more likely to be detected, which will naturally reduce operators' misses rate. Meanwhile, different failure modes also had significant impact on misses rate. Misses rate under false alarms condition was significantly lower than misses condition. This result indicates that, when the automation system has misses errors, operators tend to have

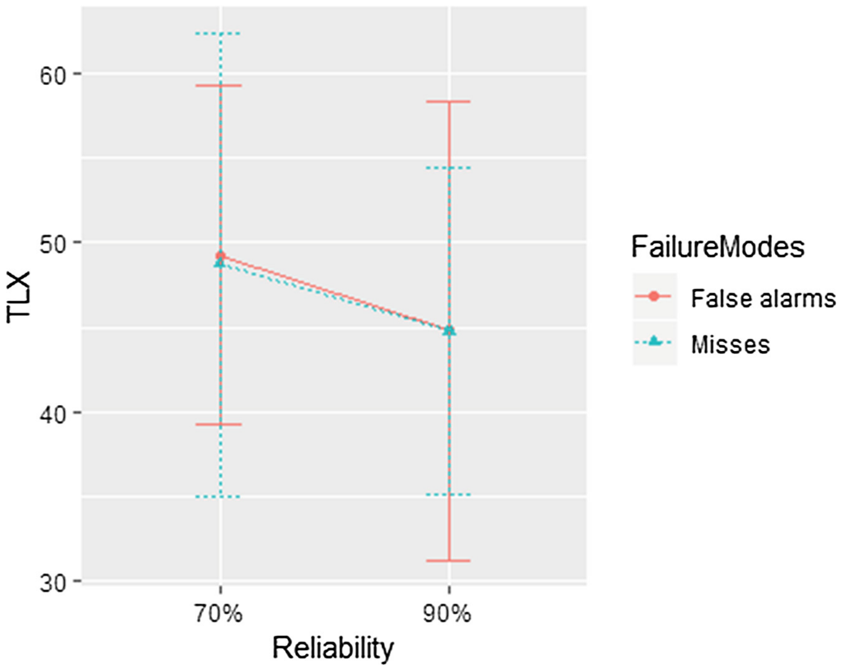


Fig. 8. NASA-TLX scores

higher misses rates. However, when the automation system becomes more "cautious" and reports every potential danger, the operators also have lower misses rates. So, in order to improve the security, besides improving the reliability of automation systems, the design of the system should also be more "critical" to ensure every potential danger is detected.

5.2 Efficiency Related Items: False Alarms and RT

Variables related to efficiency mainly include false alarms rate and reaction time. It is obvious that the longer reaction time is, the lower is the efficiency. And when false alarms occur in real situations, operators need to pause the conveyor and open the baggage to check, so this will also influence the efficiency. From the result of ANOVA, the reaction time in target-present cases decreased as the reliability increased. What's more, the failure mode of the system had no significant impact on reaction time and marginal significant impact on operators' false alarm rate. In other words, when the system is more "critical", operators' performance has a similar trend (more likely to have false alarms) to some extent, just marginally. This result indicates that, when we design the system more critically aiming to improve security, the negative impact on efficiency is not that large, still acceptable. So, in the trade-off of security and efficiency, we can lay more weight on security.

6 Conclusion

This study aims to investigate the impacts of automation reliability and failure modes on operators' performance in security screening. Results indicate that higher reliability and critical system design (false alarms) can improve security, while the latter has some negative impacts on efficiency. Meanwhile, in the trade-off of security and efficiency, more weight should be laid on security.

References

Singh, A.L., Tiwari, T., Singh, I.L.: Effects of automation reliability and training on automation induced complacency and perceived mental workload. J. Indian Acad. Appl. Psychol. **35**, 9–22 (2009)

Chun, M.M., Wolfe, J.M.: Just say no: how are visual searches terminated when there is no target present? Cogn. Psychol. **30**(1), 39–78 (1996)

Dixon, S.R., Wickens, C.D.: Automation reliability in unmanned aerial vehicle control: a reliance-compliance model of automation dependence in high workload. Hum. Factors **48**(3), 474 (2006)

Fiondella, L., Gokhale, S.S., Lownes, N., Accorsi, M.: Security and performance analysis of a passenger screening checkpoint for mass-transit systems. In: Homeland Security, vol. 43, pp. 312–318. IEEE (2013)

Hillesheim, A.J., Rusnock, C.F.: Predicting the effects of automation reliability rates on human-automation team performance. In: Winter Simulation Conference, pp. 1802–1813. IEEE (2017)

Liang, K.J., et al.: Automatic threat recognition of prohibited items at aviation checkpoint with X-ray imaging: a deep learning approach. In: Anomaly Detection and Imaging with X-Rays (ADIX) III. Proceedings of SPIE, p. 1063203, 27 April 2018

Mery, D., Svec, E., Arias, M.: Object recognition in baggage inspection using adaptive sparse representations of X-ray images. In: Bräunl, T., McCane, B., Rivera, M., Yu, X. (eds.) PSIVT 2015. LNCS, vol. 9431, pp. 709–720. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-29451-3_56

Rice, S., Mccarley, J.S.: Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. J. Exp. Psychol.: Appl. **17**(4), 320–331 (2011)

Rovira, E., Mcgarry, K., Parasuraman, R.: Effects of imperfect automation on decision making in a simulated command and control task. Hum. Factors **49**(1), 76–87 (2007)

Donaldson, T.S.: Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. J. Am. Stat. Assoc. **63**, 660–676 (1968)

Lunney, G.H.: Using analysis of variance with a dichotomous dependent variable: an empirical study. J. Educ. Meas. **7**(4), 263–269 (1970)

How Task Level Factors Influence Controllers' Backup Behavior: The Mediating Role of Perceived Legitimacy and Anticipated Workload

Saisai Yu[1,2], Jingyu Zhang[1,2(✉)], and Xiaotian E[1,2]

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

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

Abstract. The volume of air traffic has increased considerably in recent years, and the task load of air traffic controllers (ATCos) is reaching a new high. Since the mental workload of ATCos is linked with both safety and efficiency of aviation, both researchers and practitioners are seeking novel methods to prevent overload. In this research, we adopted a new approach to understanding the workload management of ATCos by investigating how they made backup decisions. The aim of the research is to investigate the forms and mechanisms behind cross-sector backup of ATCos. Based on literature review and expert interview, we identified three task-level variables (task load of providers, task load of requestors, and close-landing demands of the to-be-hand-over aircraft) and two mediating variables (workload of participants and the perceived legitimacy of backup requests) that may influence controllers backup decisions in parallel runway operations, a typical and important form of ATCos cooperation. To validate this model, we conducted two studies. We invited licensed controllers to perform simulated final approach scenarios on a medium-fidelity ATC simulation platform. They had to decide whether to accept a hand-over request made by a controller working in the neighboring sector. In Study 1, three task-level variables (task load of participants, task load of requesters, and the close-landing demands of the to-be-hand-over aircraft) were manipulated, and two mediating variables (workload of participants and evaluations of the legitimacy of backup requests) were measured. HLM analysis firstly showed that task-level variables all significantly predicted backup decisions. Controllers were more willing to accept the request when they were under low pressure, when their colleagues were at higher pressure and when the aircraft had a close-landing demand. As for two mediating variables, participants perceived legitimacy of requests mediated the relationships between task-level variables and back-up decisions. However, the perceived current workload of participants did not mediate the impact of task variables on backup outcomes as expected. We proposed that it was the anticipated workload of controllers, not the perceived current workload of controllers, that played the mediating role. In Study 2, the anticipated workload of participants was measured in addition to the other two mediating variables. HLM analyses suggested that both perceived legitimacy and anticipated workload were mediators between task-level variables and back

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 150–164, 2019.
https://doi.org/10.1007/978-3-030-22507-0_12

up decisions. In conclusion, this study identified several key psychological factors influencing ATCos cross-sector backup behaviors for the first time.

Keywords: Air traffic control · Parallel final approach · Cross-sector cooperation · Backup behaviors · Individual difference

1 Introduction

With the rapid development of the aviation industry, the current airspace is facing a saturation problem. Meanwhile, the shortage of air traffic controllers (ATCos) makes the situation even worse by casting too much workload to incumbent ATCos [23]. Since ATCos workload is an important constraint of aviation safety and efficiency, many studies had examined the factors that can predict or reduce controller's workload at an individual level [26–28, 35, 41, 46]. However, there is another way for controllers to manage their workload: receiving and providing backup to their teammates.

Backup behaviors are generally defined as helping other team members to perform their roles [8, 33]. Since it can compensate and redistribute the unbalanced resources at the team level, backup has been considered as one of the essential aspects of teamwork [31]. Studies have found evidence that backup behavior can improve team effectiveness [10, 36]. However, few studies have investigated the backup behavior in ATCos (for an exception study which used survey method, sees [49]).

In this study, we will focus on backup behaviors in typical and important cooperation: parallel runway operation (PRO) [11, 44]. PRO is common for ATCos who manage sectors around any large airport which has multiple runways. In performing a PRO, two approach controllers issue orders to pilots in their sectors to guide their queueing, landing and taking off using the runway in their sectors. At the same time, they have to pay attention to each other's sector because the wake stream caused by landing/taking off using one runway may influence the other [22]. In this operation, controllers often offer and provide backup. Typically, one controller (the requestor) may request a handover of a certain aircraft to the neighboring sector, and the controller of that adjacent sector (the provider) need to ponder whether to accept such handover request. Figure 1 illustrates the situation of a typical PRO and a backup situation. Since it is more important to understand why certain backup can be accepted, we would first review the factors that may influence the backup decisions of the provider.

1.1 Task Level Factors in Predicting Providers Backup Decision

The most important task-related variables that can influence providers backup behavior is the task-load of both the requestor and the provider [3, 39, 40]. Obviously, the providers are less likely to provide any backup if their task load is already very high [40]. From the perspective of cognitive resources theories [25, 37, 51], if the task requirement is beyond their cognitive capacity, the providers are not able to provide any back up because it may consume their resource to deal with their task and undermine their own task performance. Lots of studies showed that it is critical for the controllers to avoid overload since aviation safety is in the first place [7, 27, 28, 41, 45, 47]. Therefore,

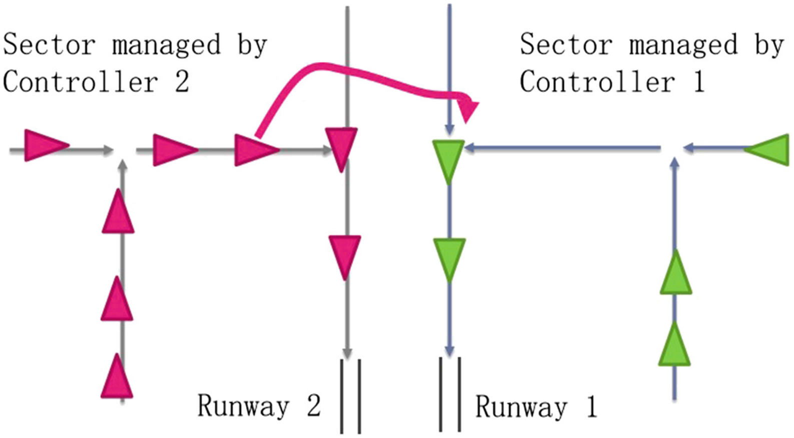


Fig. 1. The situation of a typical PRO and a backup situation

we speculate that the task load of backup providers is an important predictor of the backup decision. In ATC, the task load of a controller can be evaluated by the traffic complexity metrics which is often quantified by the number of aircraft in any given sector [18, 29, 32]. Accordingly, we proposed our first hypothesis:

H1. *The task load of backup providers, as quantified by the number of aircraft in the sector, is negatively related to the possibility of backup behaviors.*

It was also found that the providers were more likely to provide backup if the task load of the requestor is higher [3, 39, 40]. This is because the request must be considered reasonable by the provider to be accepted. If someone who makes a request for help but does not have much work to do, the provider may question his/her motivation and treat it as a form of social loafing [16, 34]. In this way, the task-load of the backup requestor is a signal showing the requestor has a genuine need, and that makes the provider willing to offer help. Accordingly, we proposed our second hypothesis:

H2. *The task load of backup requestors, also quantified by the number of aircraft in their sectors, is positively related to the possibility of backup behaviors.*

While the two abovementioned forms of task load are important in previous studies [3], we would discuss a new and unique factor that may have a great impact in influencing controllers' backup decision: the close-landing demands of the to-be-hand-over aircraft. We call that an aircraft has a close-landing demand if handing over such an aircraft to another sector can reduce the ground taxiing distance. This is because generally most commercial planes need to port on their own company's gates located on one side of the airport. However, these gates might be far away from the runway they landed if no transfer is made. For example, let us think company A's gates are located near the Runway 1 as shown in Fig. 1, then if a plane of company A is coming from the west, it should be managed by controller 2 and uses runway 2 to land. Therefore, it has to taxi a long distance to reach its gate. However, if such an aircraft is handed over from controller 2 to controller 1, the taxiing distance will be greatly shortened which may, in turn, reducing the waiting time of passengers and the costs of airline companies. In this situation, such a plane has a closing land demand. To note, it is not a formal requirement that all controllers must accept such kinds of aircraft. However, controllers may find it to be a good favor and may accept this kind of request in certain circumstances [13]. Therefore, accordingly, we proposed our third hypotheses:

H3. *The close-landing demands of to-be-hand-over aircraft are positively related to the possibility of backup behaviors. If an incoming aircraft has close-landing demand, backup providers are more willing to accept the backup request.*

1.2 The Mediating Variables

To better reveal the inner mechanism behind the backup decision-making process, we further listed two variables that may mediate the impact of three task-level variables previously mentioned.

The first mediating variable is mental workload. Whereas previous studies manipulated the actual task load by changing, for example, the number of targets, how the operators did perceive has not been measured. Mental workload is an important measure used in human factor studies which reflects the surplus of the operators' capacity as meeting the task demand. Often measure using subjective measures, the mental workload can provide additional explanatory power beyond actual task load since it also takes the capacity of the operator into account. The same level of task load (e.g., four aircraft in a sector) can result in different levels of mental workload upon controllers with different amount of cognitive resources. For example, an experienced controller may find it very easy to handle (low mental workload) while a novice may find it extremely hard (high mental workload). In this way, since mental workload reflects the unused resources to operating an addition task (i.e., backup), it may mediate the influence of providers task load on their backup behaviors.

H4. *The provider's mental workload mediates the influence of providers task load, as quantified by aircraft number, on backup behaviors.*

The second mediating variable is perceived legitimacy. Whereas mental workload reflects whether a controller is ABLE to provide backup, perceived legitimacy reflects whether a controller is WILLING to do so. Although such a concept has been raised by previous studies [3, 39, 40], it was only manipulated by the imbalance of task load rather than being directly measured. There are two problems to use such kind of manipulation. First, it is not known whether the backup behaviors are made due to a genuine feeling of legitimacy or fairness or just simply because they do not have the resource to do so. Second, it precludes other factors beyond task load imbalance that may also result in the feeling of legitimacy. For example, in our study, the existence of close-landing demand may also increase the legitimacy of the request since that is a good thing to do. As a result, in this study, we intended to measure this variable directly to see whether the provider has a feeling of legitimacy and whether such a feeling can mediate the influence of task-level factors (i.e., requestors task load and close-landing demand) and backup behaviors.

H5. *The perceived legitimacy of backup requests mediates the influence of requestors task load and close-landing demand on the possibility of backup behaviors.*

2 Study 1

In order to test our hypotheses, we manipulated the three task-level variables and measured the two mediating variables and conducted hierarchical linear modeling (HLM) to analyze the data.

2.1 Method

Participants. In total, 22 licensed professional air traffic controllers participated in this experiment. Their ages ranged from 24 to 48 years (M = 29.11, SD = 6.41) and ATC experience ranged from 2 to 20 years (M = 5.94, SD = 4.74). Due to equipment failure, only 18 were available for analysis. All participants were paid 150 RMB after completing this experiment. All participation was voluntarily and anonymously.

The Parallel Runway Operation Task and Scenarios. ATC-Simulator, a medium fidelity ATC simulation platform [12, 52–54], is used to simulate the parallel runway operation. In each scenario, there were two final approach sectors each of which contained a runway. All participants managed the sector on the right side of the screen, while the adjacent side on the left side was operated by another hypothetical ATCo performing the pre-planned operation. The sketch map is shown in Fig. 2. Each plane had a label showing its call sign, direction, altitude, course, and speed. Participants

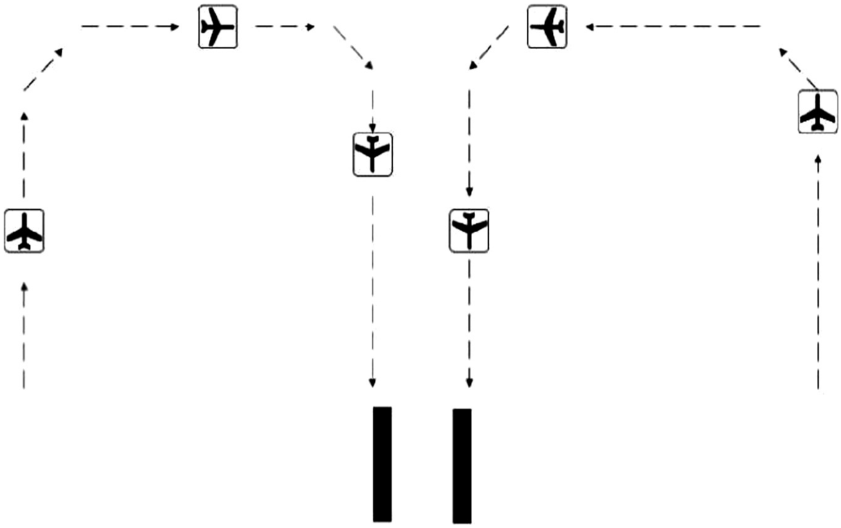


Fig. 2. Task parallel final approach interface

could use two supportive tools. One is a scale of 10 nm * 20 nm located in the lower left corner of the screen. The another is a distance/time calculation tool to get the distance and angle from the former point to the latter point and the angle, time and distance of aircraft flying from the current position to the point at current speed. The participants can see the conditions in both sectors, but they can only issue orders to aircraft in their own sectors.

At the beginning of each scenario, multiple aircraft appeared in both sectors, and the participants are required to constantly monitor and adjust the speed and altitude of the aircraft to fulfill the following three requirements: (1) do not violate the minimum separation standard (5 nautical miles level, 1000 ft vertical); (2) keep the speed of aircraft less than 200 knots and the altitude of aircraft less than 3000 ft when entering the Final; (3) keep a 5 min time interval between aircraft. This period lasted for 40 s, during which the participants were requested to make necessary interventions. After that, a series of dialog boxes popped up to collect the ratings of mental workload using the six items of NASA-TLX [20]. The average score of the six items was used as the mental workload ratings.

After answering these questions, the task was frozen, and the participants were told that the colleague of the neighboring sector wanted to hand over an aircraft due to certain reasons, and asked how they would think and respond to this request. In this process, participants could see their current flight situation and the to-be-handed-over aircraft, but they cannot make any interventions. Perceived legitimacy was measured by using an 8-point item "how legitimate do you think the request is?" (1 representing very low and eight very high). Backup willingness was measured by an 8-point item "how is your willingness to accept the handover aircraft" (1 representing very low and eight very high). The backup decision was measured by a dichotomous force choice. "Do you accept or reject the aircraft?" (0 representing rejection and one acceptance). When participants completed all the questions, they would enter the next scene.

The whole task had 38 scenarios, the first 6 were practice scenarios, and the remaining 32 were for formal experiments. The experiment used a 2 (participants task-load: low/high) * 2 (requestors task-load: low/high) * 2 (close-landing demand: have/no) within-subjects design, resulting in 8 different conditions and each condition contained 4 different scenarios.

The close-landing demand was manipulated by the call sign of the aircraft to be handed over. In the situation with a close-landing demand, the aircraft to be handed over will board to a gate near the participant's sector; in the situation without a close-landing demand, the aircraft to be handed over will board to a gate near the colleague's sector.

The task load of participants was manipulated by the number of aircraft in their sector. There were four aircraft in the low task load condition and 10 in the high task load condition. The task load of backup requestors was manipulated by the number of aircraft in the left sector. There were two aircraft in the low task load condition and 8 in the high task load condition.

Experimental Process. Upon arrival, the experimenter briefed all participants on the process of the experiment, and the participants signed the informed consent. Next, the participants were asked to remember several call signs of the airline companies and the locations of their gates (close to their sector or their colleague's sector). They were tested then to ensure all that information was accurately remembered. Afterward, they completed the six practice scenarios and 32 formal scenarios on a computer with a 23-in. monitor, which lasted about one hour. Participants reported demographics (sex, age, work experience) afterward and they were paid, thanked and debriefed.

2.2 Results

Basic Analysis. Table 1 provides the means (M), standard deviations (SD) the correlations of all variables. From the table, we can see that both backup wiliness and decisions were significantly correlated with the three task-level variables and the two mediating variables.

Table 1. Means, standard errors and zero-order correlations of all variables in study 1 (n = 18)

	M (SD)	1	2	3	4	5	6	7
1. Close-landing demand	.50 (.50)							
2. Task load of the provider	7.00 (3.01)	-						
3. Task load of the requestor	5.00 (3.01)	-	-					
4. Mental workload	2.28 (1.28)	−.01	.60**	.03				
5. Perceived legitimacy	3.58 (2.31)	.10*	−.58**	.24**	−.37**			
6. Backup willingness	3.73 (2.42)	.09*	−.62**	.17**	−.38**	.89**		
7. Backup decision	.43 (.45)	.09*	−.66**	.11**	−.37**	.63**	.68**	
8. Job experience	5.94 (4.74)	-	-	-	−.04	.07	.03	.05

Multilevel Regression Modeling. In order to test our hypotheses, we conducted multilevel regression modeling using HLM 6.08 software to analyze the data [43]. In performing the multilevel modeling, a null model with no predictors at both levels was built to test whether the data is suitable for multilevel analysis. The intra-class correlation (ICC) of backup willingness and backup behavior were 10.24% and 10.88%, respectively, suggesting the necessity to use the modeling approach [38].

Regression Models in Predicting Backup Willingness. Using backup willingness as the dependent variable, we constructed two nested models. In model 1, we put three task-level variables, close-landing demand, task-load of the provider, and task-load of the requestors, into the model. Job experience was also entered into the model as a control variable.

The results showed that the close-landing demand (β = .503, p < .01), task-load of providers (β = −.576, p < .01) and workload of requestors (β = .156, p < .01) all significantly predicted the backup willingness (see Table 2 for details). Specifically, participants were more willing to accept the aircraft if it had a close-landing demand when they own task-load was low and when requestor's task load was high. Therefore, H1 to H3 were all confirmed.

Table 2. HLM in Model 1

	Backup willingness		Backup decision (1 backup, 0 refuse)	
	Model 1	Model 2	Model 3	Model 4
Intercept	3.724^{**} (.29)	3.724^{**} (.29)	.503 (.63)	.542 (.31)
Individual level variable (N = 18)				
1. Job experience	.006 (.06)	.006 (.06)	.034 (.13)	.034 (.06)
Task level variable (N = 576)				
2. Close-landing demand	$.503^{**}$ (.14)	.076 (.09)	$.711^{**}$ (.22)	$.545^{*}$ (.16)
3. Provider's task load	$-.576^{**}$ (.02)	$-.182^{**}$ (.02)	$-.590^{**}$ (.04)	$-.377^{**}$ (.05)
4. Requestor's task load	$.156^{**}$ (.02)	−.011 (.01)	$.147^{**}$ (.03)	.059 (.04)
5. Perceived legitimacy		$.778^{**}$ (.02)		$.426^{**}$ (.07)
6. Mental workload		.023 (.06)		−.194 (.17)

Note: *p < .05, **p < .01

In model 2, the perceived legitimacy and mental workload were added into the model. The results showed that perceived legitimacy was a significant predictor of backup willingness *(β* = .778, p < .01), while mental workload had no significant influence (*β* = .023, *n.s.*). At the same time, the effects of the three task-level variables were all reduced. The close-landing demand was dropped from 0.053 to 0.076 (*β* = .076, *n.s.*), the task-load of the provider was dropped from 0.576 to 0.182 (*β* = −.011, *n.s.*) and task-load of the requestor was drop from 0.156 to 0.011 (*β* = −.182, p < .01). It means that perceived legitimacy mediated the influence of all three task-level variables on the backup willingness. Therefore, H4 was not confirmed, but H5 was confirmed.

Regression Models in Predicting Back-Up Decision-Making. Using backup decision-making as the dependent variable, we constructed two nested models which were similar to multilevel regression modeling using HLM 6.08 software to analyze the data [21]. In model 3, we put three task-level variables, close-landing demand, task-load of the provider, and task-load of the requestors, into the model. Job experience was also entered into the model as a control variable. Job experience was also entered into the model as a control variable.

The results showed that the close-landing demand (*β* = .711, p < .01), task-load (*β* = −.590, *p* < .01) and requestors workload (*β* = .147, p < .01) all significantly predicted the backup decision-making. Specifically, participants were more willing to accept the aircraft if it had a close-landing demand, when their task-load was low and when requestors' task load was high. Therefore, H1 to H3 were all confirmed.

In model 4, the perceived legitimacy and mental workload were added into the model. The results showed that perceived legitimacy was a significant predictor of backup decision-making (*β* = .426, p < .01), while mental workload had no significant influence (*β* = .194, *n.s.*). At the same time, the effects of the three task-level variables were all reduced. The task-load of the requestor was a drop from 0.147 to 0.059 (*β* = −.182, p < .01) which was no longer significantly predicted the backup decision-making. The close-landing demand (*β* = .545*,* *p* < .05) and the task-load of the provider ($\beta = .377, p < .01$) was still predicted backup decision-making, but the effects were weakened. The perceived legitimacy mediated the influence of all three task-level variables on backup decision-making. Therefore, H4 was not confirmed, but H5 was confirmed.

2.3 Discussion of Study 1

The results of Study 1 fully confirmed H1 to H3 suggesting the functions of the task-level variables were in agreement with our expectations. Also, it confirms H5 suggesting perceived legitimacy plays an important mediating role. However, H4 was not supported suggesting mental workload could neither predict backup willingness nor decision. One possible reason is that when making a decision of the future, it is not the current mental workload of controllers that matters. When deciding whether to accept a backup request that happens in the near future, the controllers may need to evaluate whether he/she will have available mental resource in that period rather than contemplating how he/she experienced in the past (what the mental workload variable measured in study 1). Literature in ATC decision making also pointed out the importance of making future anticipations, such as trajectories [12] or mental workload [26]. As a result, we suggested that it might be anticipated workload, rather than current mental workload, that would play the mediating role. Therefore, we modified H4a as:

H4a. *The anticipated mental workload of the participants is positively related to the possibility of backup behaviors in ATC.*

In order to replicate the major findings of study 1 and our modified new hypothesis, we conducted study 2.

3 Study 2

In study 2, we sought to make a replication of study and test our new hypothesis related to anticipated mental workload. In doing so, we used similar task configurations as compared to Study 1 but adding a new item measuring anticipated mental workload.

3.1 Participants

In total, 22 licensed professional ATCos participated in this experiment. Their ages ranged from 22 to 34 years (M = 26.50, SD = 3.14) and ATC experience ranged from 1 to 11 years (M = 4.18, SD = 2.99). Finally, all of the data were available for analysis. All participants were paid 150 RMB after completing this experiment. All participation was voluntarily and anonymously.

3.2 Method

All experimental settings were similar to Study 1, except for adding a new item to measure anticipated mental workload. This item was added before responding to the question about the perceived legitimacy of backup. Anticipated mental workload was measured by using an 8-point item "how much do you think it will bring you extra work if you accept the handover aircraft?" (1 representing very low and eight very high).

3.3 Results

Basic Analysis. Table 3 provides the means (M), standard deviations (SD) the correlations of all variables. From the table, we can see that both backup willingness and decisions were significantly correlated with the three task-level variables and the two mediating variables.

Table 3. Means, standard errors and zero-order correlations of all variables in study 2 (n = 22)

	M (SD)	1	2	3	4	5	6	7	8
1. Close-landing demand	.50 (.50)								
2. Task load of the provider	7.00 (3.00)	-							
3. Task load of the requestor	5.00 (3.00)	-	-						
4. Mental workload	2.24 (1.68)	.001	.60**	.02					
5. Anticipated mental workload	4.41 (2.90)	−.001	.80**	−.04	.65**				
6. Perceived legitimacy	3.47 (2.67)	.05	−.54**	.18**	−.37**	−.60**			
7. Backup willingness	3.71 (2.75)	.03	−.65**	.13**	−.48**	−.72**	.81**		
8. Backup decision	.47 (.50)	.06	−.70**	.08**	−.51**	−.76**	.62**	.75**	
9. Job experience	4.18 (2.99)	-	-	-	−.03	−.13	.02	.04	.07

Multilevel Regression Modeling. Similar to Study 1, we established the models by using backup willingness and backup behavior as the dependent variables. In the null models, the intra-class correlation (ICC) of backup willingness and backup behavior were 9.81% and 9.71%, respectively, suggesting the necessity to use the multilevel modeling approach.

Next, we entered Job experience and the three task-level variables, into the first step (model 1 and model 4). Similar to study 1, the result showed that all three task-level variables (close-landing demand, provider's task-load, and the requestor's task load) all had significant influences on participants' backup willingness and backup behavior (see Table 2 for details).

In the second step, the perceived legitimacy and the current mental workload were added into model 2 and model 5. Similar to study 1, the perceived legitimacy mediated the effect of task level variables on willingness ($\beta = .614$, $p < .01$) and backup behavior ($\beta = .503$, $p < .01$). Current mental workload, however, also had a significant influence over backup willingness ($\beta = -.177$, $p < .05$) and backup behavior ($\beta = -.238$, $p < .05$).

In the final step, the anticipated workload was added into model 3 and model 6. It was found that anticipated mental workload significantly predicted backup willingness ($\beta = -.275$, $p < .01$) and backup behavior ($\beta = -.411$, $p < .01$). Moreover, adding anticipated mental workload into the models significantly reduced the effects of current mental workload on the backup willingness ($\beta = -.114$, *n.s.*) and backup behavior ($\beta = -.124$, *n.s.*). Therefore, the anticipated mental workload was a more proximal predictor of back up behavior, which played the mediating role between task-level variables and backup behavior. Therefore, H4a was confirmed.

3.4 Discussion of Study 2

As did in study 1, Study 2 further confirmed H1 to H3 suggesting the three task-level variables were important predictors of backup willingness and decision. Also, perceived legitimacy was found to play the mediating role again. Moreover, study 2 also supports H4a, suggesting it is an anticipated mental workload that plays the mediating role in making a backup decision (Table 4).

Table 4. Hierarchical linear model in experiment 2

Parameter	Backup willingness			Backup decision (1 backup, 0 refuse)		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	3.710** (.20)	3.710** (.20)	3.710** (.20)	.211 (.26)	.417 (.24)	.493 (.26)
Individual-level variable (N = 22)						
1. Job experience	.041 (.06)	.041 (.06)	.041 (.06)	.116 (.08)	.123 (.08)	.104 (.08)
Task-level variable (N = 704)						
2. Close-landing demand	.164 (.16)	−.026 (.07)	−.010 (.07)	.556* (.24)	.527* (.20)	.601* (.23)
3. Task load of provider	−.601** (.04)	−.242** (.06)	−.076* (.03)	−.615** (.05)	−.414** (.06)	−.198** (.05)
4. Task load of requestor	.125** (.03)	.026 (.02)	.023 (.01)	.108* (.05)	.026 (.03)	.020 (.02)
5. Perceived legitimacy		.614** (.07)	.556** (.08)		.503** (.08)	.465** (.09)
6. Current mental workload		−.177* (.08)	−.114 (.08)		−.238* (.11)	−.124 (.10)
7. Anticipated mental workload			−.275**(.07)			−.411** (.07)

notes: *p < .05, **p < .01

4 Overall Discussion

In our research, we attempted to explore the mechanisms underline the backup behaviors of ATCos from the viewpoint of backup providers. In testing our hypotheses, we conducted two studies by recruiting professional ATCos in performing a parallel runway operation task.

Across two studies, we found that task-load distribution was an important predictor of controllers' backup decision making: controllers are more likely to provide help when their own task-load is low and when their colleague's task-load is high. These results, fully confirming H1 and H2, are consistent with previous studies in general areas of cooperation [3, 39, 40]. This finding provides a new way to look at the critical issue of controllers' mental workload. Although overload may occur at some time for some controllers, as long as there are other underload controllers at the same time, backup behaviors can help adjust they workload distribution in increasing the overall performance of aviation efficiency and safety.

Besides, the close-landing demand was found to affect the willingness and final decision of backup. Specifically, if the aircraft had a close-landing demand, controllers are more likely to accept the hand-over of this kind of aircraft. Confirming H3, this factor was first identified in the literature of ATC and team backup behaviors. To note, although accepting aircraft with a close-landing demand can save the waiting time of the crew and passengers, it is neither required nor beneficial for the personal interests of the controllers. Indeed, such a kind of behavior can lead to extra effort on the side of the providers. As a result, this phenomenon can be seen as an altruistic and prosocial behavior that is beyond team cooperation. Future studies may take further steps to see other similar conditions that may contribute to controllers' backup behaviors, in the scope of organizational citizenship or customer service literature.

In addition, both studies found that perceived legitimacy played an important mediating role (H5). This finding corroborates previous theoretical argument that a legitimacy evaluation process is involved in making the backup decision making [3, 39, 40]. However, our study is the first to measure this construct empirically. Since there are still a large number of individual differences behind this perception, future studies may further explore how people form different legitimacy evaluation upon similar targets and situations.

One of the most interesting findings of the present study was that it was the anticipated mental workload in the future that plays a more important role in forging controllers' backup decision, as compared to their current mental workload. However, a lot of ATC researches mainly focus on the measurement and the function of the latter. According to our analysis, the future workload may have a greater influence on the controller's strategy choice and behaviors. This finding is inconsistent with some new progress in the domain of general decision-making [1, 4, 48] and some studies focusing on mental workload management in the area of ATC [26, 47]. Future studies may benefit from exploring the formation and function of this variable. For example, how to help controllers to form an accurate evaluation of their future workload? How to help them make better decisions using this information?

Before concluding, it is important to discuss some potential limitations of the study. First, this study only utilized a relatively small sized sample; however, this is a common practice for studies based on professional controllers and other similar occupational groups (e.g., [5, 17, 26, 30, 42, 46, 47, 50, 53]). However, we ensured the validity of the study by using two experiments to survey a repeated verification. This procedure, while minimizing the potential confounding variables such as interpersonal relationships and social ranking, can offer a better understanding of the task-level properties. Future studies may benefit from including more "social" variables into the exploration, such as familiarity and team personalities, which, however, requires the researchers to have a much larger pool of participant.

5 Conclusion

In this paper, we focused on the cross-sector backup and recruited professional controllers to conduct dynamic parallel final approach missions. We found that the ATCos were more likely to provide help to their colleagues if the request is made when their task load is low, when their colleagues' load is high and when the request can benefit the crew and passengers. Also, the effects of these variables were mediated by two psychological variables: the perceived legitimacy and the anticipation of future mental workload of the backup.

Acknowledgments. This research was supported by the National Key Research and Development Plan [grant number: 2016YFB1001203] and the Natural Scientific Foundation of China [31671148]. We thank Ms. Yazhe Li, Mr. Chen Zhang, and Dr. Jiashu Zhou in their help during the experiments.

References

1. Botvinick, M.M., Rosen, Z.B.: Anticipation of cognitive demand during decision-making. Psychol. Res. PRPF **73**(6), 835–842 (2009)
2. Byron, K.: Carrying too heavy a load? The communication and miscommunication of emotion by email. Acad. Manag. Rev. **33**(2), 309–327 (2008)
3. 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)
4. Botvinick, M.M.: Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. Cogn. Affect. Behav. Neurosci. **7**(4), 356–366 (2007)
5. Boag, C., Neal, A., Loft, S., Halford, G.S.: An analysis of relational complexity in an air traffic control conflict detection task. Ergonomics **49**(14), 1508–1526 (2006)
6. Byron, K., Baldridge, D.C.: Toward a model of nonverbal cues and emotion in email. Acad. Manag. Annu. Meet. Proc. **2005**(1), B1–B6 (2005)
7. Bisseret, A.: Application of signal detection theory to decision making in supervisory control: the effect of the operators' experience. Ergonomics **24**(2), 81–94 (1981)
8. Dickinson, T.L., McIntyre, R.M.: A conceptual framework for teamwork measurement. In: Brannick, M.T., Salas, E., Prince, C. (eds.) Team Performance Assessment and Measurement: Theory, Methods, and Applications, pp. 19–43. Erlbaum, Mahwah (1997)
9. Daft, R.L., Lengel, R.H.: Organizational information requirements, media richness and structural design. Manag. Sci. **32**(5), 554–571 (1986)
10. De Dreu, C.K.W.: Cooperative outcome interdependence, task reflexivity, and team effectiveness: a motivated information processing perspective. J. Appl. Psychol. **92**(3), 628–638 (2007)
11. Domino, D.A., Tuomey, D., Mundra, A., Smith, A., Stassen, H.P.: Air ground collaboration through delegated separation: results of simulations for arrivals to closely spaced parallel runways. In: 2011 Integrated Communications, Navigation, and Surveillance Conference Proceedings (2011)
12. 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)
13. E, X.: Masters thesis of the Institute of Psychology. Chinese Academy of Sciences (2018)

14. Ekman, P.: Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. W.W Norton Press, New York (2009)
15. Ekman, P., Friesen, W.V., Scherer, K.: Body movement and voice pitch in deceptive interaction. Semiotica **16**, 23–27 (1976)
16. Erez, M., Somech, A.: Is group productivity loss the rule or the exception? Effects of culture and group-based motivation. Acad. Manag. J. **39**(6), 1513–1537 (1996)
17. Fothergill, S., Neal, A.: The effect of workload on conflict decision making strategies in air traffic control. Hum. Factors Ergon. Soc. Annu. Meet. Proc. **1**(1), 39–43 (2008)
18. Gianazza, D.: Forecasting workload and airspace configuration with neural networks and tree search methods. Artif. Intell. **174**(7–8), 530–549 (2010)
19. Hancock, J.T., Landrigan, C., Silver, C.: Expressing emotion in text-based communication. In: Proceedings of the CHI 2007 Conference on Human Factors in Computing Systems, pp. 929–932. Association for Computing Machinery Press, New York (2007)
20. 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.) Human Mental Workload, pp. 139–183. North Holland Press, Amsterdam (1988)
21. Hox, J.: Multilevel Analysis: Techniques and Applications. Lawrence Erlbaum Associates, Mahwah (2002)
22. Manual on Simultaneous Operations on Parallel or Near-Parallel Instrument Runways (SOIR), 1st edn. International Civil Aviation Organization, 9643-AN/941 (2004)
23. Global Air Transport Outlook to 2030 and trends to 2040 (No. Cir 333, AT, 190). ICAO, Montreal, Canada (2013)
24. Kallus, K.W., Van Damme, D., Dittman, A.: Integrated job and task analysis of air traffic controllers: Phase 2. Task analysis of en-route controllers (European Air Traffic Management Programme Rep. No. HUM.ET1.ST01.1000-REP-04). EUROCONTROL, Brussels, Belgium (1999)
25. Kahneman, D.: Attention and Effort. Prentice-Hall, Englewood Cliffs (1973)
26. 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)
27. 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)
28. Loft, S., Humphreys, M., Neal, A.: Prospective memory in air traffic control. In: Australian Aviation Psychology Symposium, vol. 1, pp. 287–294. Ashgate Publishing Company (2003)
29. 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 (1998)
30. Metzger, U., Parasuraman, R.: The role of the air traffic controller in future air traffic management: an empirical study of active control versus passive monitoring. Hum. Factors **43**(4), 519–528 (2001)
31. McIntyre, R.M., Salas, E.: Measuring and managing for team performance: lessons from complex environments. In: Guzzo, R.A., Salas, E. (eds.) Team Effectiveness and Decision-Making in Organizations, pp. 9–45. Jossey-Bass, San Francisco (1995)
32. 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. CTA Inc., Mckee City (1995)
33. Morgan Jr., B.B., Glickman, A.S., Woodard, E.A., Blaiwes, A., Salas, E.: Measurement of team behaviors in a Navy environment (NTSC Report No. 86-014). Naval Training System Center, Orlando (1986)

34. Murphy, S.M., Wayne, S.J., Liden, R.C., Erdogan, B.: Understanding social loafing: the role of justice perceptions and exchange relationships. Hum. Relat. **56**(1), 61–84 (2003)
35. 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)
36. Neuman, G.A., Wright, J.: Team effectiveness: beyond skills and cognitive ability. J. Appl. Psychol. **84**(3), 376–389 (1999)
37. Norman, D.A., Bobrow, D.G.: On data-limited and resource-limited processes. Cogn. Psychol. **7**(1), 44–64 (1975)
38. Peugh, J.L.: A practical guide to multilevel modeling. J. Sch. Psychol. **48**(1), 85–112 (2010)
39. 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)
40. 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)
41. 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)
42. Rantanen, E.M., Nunes, A.: Hierarchical conflict detection in air traffic control. Int. J. Aviat. Psychol. **15**(4), 339–362 (2005)
43. Raudenbush, S.W., Bryk, A.S.: Hierarchical Linear Models: Applications and Data Analysis Methods, vol. 1. Sage Publications, Thousand Oaks (2002)
44. Robeson, I., Clarke, J.P.: A departure regulator for closely spaced parallel runways. In: 29th Digital Avionics Systems Conference (2010)
45. 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)
46. 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)
47. Sperandio, J.C.: Variation of operator's strategies and regulating effects on workload. Ergonomics **14**(5), 571–577 (1971)
48. Solomon, R.L.: The influence of work on behavior. Psychol. Bull. **45**(1), 1–40 (1948)
49. Smith-Jentsch, K.A., Kraiger, K., Cannon-Bowers, J.A., Salas, E.: Do familiar teammates request and accept more backup? Transactive memory in air traffic control. Hum. Factors **51** (2), 181–192 (2009)
50. Vuckovic, A., Kwantes, P., Neal, A.: A dynamic model of decision making in ATC: adaptation of criterion across angle and time. Hum. Factors Ergon. Soc. Annu. Meet. Proc. **55**(1), 330–334 (2011)
51. Wickens, C.M., Toplak, M.E., Wiesenthal, D.L.: Cognitive failures as predictors of driving errors, lapses, and violations. Accid. Anal. Prev. **40**(3), 1223–1233 (2008)
52. 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)
53. Zhang, J., Du, F.: Relational complexity network and air traffic controllers' workload and performance. In: Harris, D. (ed.) EPCE 2015. LNCS, vol. 9174, pp. 513–522. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20373-7_49
54. Zhang, J., Ren, J., Wu, C.: Modeling air traffic controllers decision making processes with relational complexity network. In: Presented at the 17th International IEEE Conference on Intelligent Transportation Systems, Qingdao, China (2014)

Visual Cognition

A Visual Cognition Test-Based Study on the Choice Blindness Persistence: Impacts of Positive Emotion and Picture Similarity

Huayan Huangfu[1,2], Yi Lu[1(✉)], and Shan Fu[1]

[1] Department of Automation, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
luyil@sjtu.edu.cn
[2] Counseling and Support Services, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

Abstract. Choice blindness is a common psychological phenomenon in many selection tasks. It always presents as a person's failure to notice a mismatch between one's preference and task decision. An important affecting factor of choice blindness is emotion. We studied the impacts of positive emotion and different similarity on the choice blindness based on visual cognitive tests. We designed a 2 (high and low similarity pictures) × 2 (positive emotion and neutral emotion) mixed design. 20 pairs of scenery pictures with different similarities were used as materials and 80 adults were recruited as participants. The results verified the existence of choice blindness in both positive emotion group and neutral emotion group. It was found that significant difference existed in the perception of false feedback between the positive emotion group and neutral emotion group. The female participants in the positive emotion group was more easily perceived the false feedback of low similarity scenery pictures than those female participants in the neutral emotion. In other words, positive emotion had an effect on the choice blindness of the female when they faced the low similarity scenery pictures.

Keywords: Choice blindness · Positive emotion · Picture similarity · Gender difference · Interaction effect

1 Introduction

When people make choices in life, they always think they know the reasons for their choices and can easily detect the mismatch between their choices and the results. But this is a wrong thinking about our reliability of decision-making. More and more studies have found that people may not know what their real choice is. Choice blindness is robust, repeatable and dramatic psychological phenomenon. Choice blindness is ubiquitous in choice preference and has important influence on individual decision. In theory, the study about the impact of emotion on choice blindness can provide a deeper understanding of choice blindness, deepen the study on the factors, and provide a theoretical basis for the blindness study of later researchers. In practice,

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 167–175, 2019.
https://doi.org/10.1007/978-3-030-22507-0_13

choice blindness is closely related to decision-making, which is common in consumer psychology, attitude formation, moral judgment and other fields. This study can help individuals make better decisions. Therefore, the study of choice blindness has great theoretical significance and application value.

The concept of choice blindness was first proposed and verified the existence of the phenomenon in laboratory experiments and field experiments in 2005. It means "Participants cannot find that their real choices are manipulated" [1]. In other words, it always presents as a person's failure to notice a mismatch between one's preference and task decision. The primal experiment became a classical experimental paradigm. In this experiment, participants were told that they were taking part in an experiment on the attractiveness of female faces. Then participants were presented with pairs of female face pictures and were required to make a choice between the two pictures based on attractiveness. After making a choice, the picture they chose was presented again and they were asked to explain why it was more attractive. But in fact, the participants were given false feedback in partial selection feedback. That is, the participants were presented with another photo they did not choose at first. The final results showed that up to 70% of the participants did not detect that the photos presented to them were not the original photos in the false feedback. In recent years, more and more researchers have paid attention to the mental mechanism and impact factor of choice blindness. They found that the similarity of select objects, the type of selective channel, the content and form of false feedback, personal feature of participants, the situation and so on were the impact factors. In the classic experiment, the researchers used high similarity and low similarity female face photos as the experimental materials. They found that there was no significant difference between high similarity and low similarity ones. The similarity of female face photo was not the impact factor of visual blind selection under this experimental condition [1]. But the similarity impacted on the tactile blind selection in others experiment. The more similar the select objects are, the more blind selection appeared [2, 3]. Some trails changed the materials based on the classic paradigm to explore the choice blindness. A choice blindness task used drawing pictures which were abstract painting, life painting, landscape painting and portrait painting as materials. The most participants did not notice the mismatch between their choices and the results [4]. Some researcher studied the effects of emotion on change blindness based on experiments. They found that positive emotions promoted individual awareness of change, while negative emotions hindered individual awareness of change [5]. But there is no relevant experiment about choice blindness.

The research on choice blindness is mainly concentrated in Western Europe and North America, but few in east Asia. Zhang [6] explored the psychological mechanism of choice blindness from the perspective of memory representation and recruited Chinese as participants. Although the choice blindness was persistent, the lasting time wasn't long. The memory representation theory is the cognition perspective to explain the choice blindness. Both the representation failure and the extraction failure may be the occurrence mechanism of the choice blindness, but the representation failure is likely to the basis. By studying cognitive style and sensory channel on choice blindness [7], and finding that sensory channel was an important factor of choice blindness. The participants (from china) were more likely to suffer from choice blindness under the auditory condition than the visual condition. There was also a comprehensive overview of the current studies about choice blindness was conducted and it discussed the stability of choice blindness [8].

2 Method

2.1 Participants

80 healthy adults (half male and half female, age range 18–22 years were recruited to participants in this study, and were randomly divided into two groups (each group of 40 participants, half male and half female). All participants had normal or corrected to normal vision. And they had never participated in the choice blindness experiment to ensure that they did not know the purpose of the experiment. They also had basic computer operation ability.

2.2 Material

The Chinese version of positive and negative emotion scale was made by Watson et al. and Tellege and proven cross-cultural homogeneity by Weidong Zhang. The scale presented good internal consistency (0.87), reliability and validity. The scale has 20 items and score by 5-point Likert scale. There are 10 items with the dimension of positive emotion and the left 10 items negative. 10 items with the dimension of positive emotion were used in this experiment as needed, we call it as the Chinese version of positive emotion scale.

Two pieces of video were used as the emotion induce materials. The home humor video which lasted 3 min and 58 s was used to induce positive emotion and the documentary about Palace Museum which lasted 3 min and 39 s was used to induce neutral emotion. A five minute soothing music was used to relax music.

20 pairs of scenery pictures with different similarities (10 pairs of high similarity pictures and 10 pairs of low similarity pictures) were used as select materials. All the scenery pictures were rated on a scale of 1 to 10 for similarity by 30 other participants in the pre-experiment. 1 meaned this pair of scenery pictures was completely different, and 10 meaned this pair of scenery pictures was completely identical. The average similarity of these high similarity pictures was 6.82 ± 0.48 (M ± SD), and the average similarity of these high similarity pictures was 2.78 ± 0.61 (M ± SD). Four of the twenty pairs were chosen as target pairs. Four of these pairs were false feedback pictures (2 pairs of high similarity pictures and 2 pairs of low similarity pictures), that is, the pictures presented to the participants was the non-preference pictures that the participants did not choose at the beginning.

2.3 Experiment Design

We designed a 2 (high and low similarity pictures) × 2 (positive emotion and neutral emotion) mixed design based on the classical selection of blind paradigm.

The independent variables were time interval (positive emotion and neutral emotion) and picture similarity (high similarity and low similarity). In independent variables, emotion was the variable between subjects and picture similarity was the variable within subjects. The dependent variables were the reaction after participants receiving the false feedback scenery pictures. 20 pairs of scenery pictures were presented. 16 pairs were veridical feedback pictures, and 4 pairs were false feedback. The order of veridical and false feedback pairs and pictures' similarity was presented by ABBA design.

2.4 Experiment Procedure

All participants completed the trail on the computer. First step, before the formal experiment begins, the participants were informed that "This is an experiment on scenic attraction. Please take a deep breath and relax. This experiment will not have adverse effect on you. Next, we will start the experiment, please follow the instructions to do every step."

Second step, the participants listened to the soothing music and finished the 10 items with the dimension of positive emotion of positive and negative emotion scale.

Third step, the participants watched video. One group watched the home humor video and the other group watched the documentary about Palace Museum. They finished again the 10 items with the dimension of positive emotion of positive and negative emotion scale.

Fourth step, we used the E-prime software programming to complete the experiment. A pair of scenery pictures were displayed on the left and right side of the computer screen. Participants were asked to choose the more attractive scenery picture between presented pairs. And the choice process was free. After chosen, the selected pictures or the unselected pictures, that is the false feedback will appear again in the center of the computer screen (Fig. 1). And participants were asked to write down the reasons of the chosen the preferred picture on the experimental record chart. The process was also no limited time.

2.5 Data Recording and Processing

We used E-prime and SPSS 20.0 software to process the results. The participants' unawareness of the false feedback pictures was scored as 0, their detections of once was scored as 1, and twice was scored as 2. We also recorded the reasons for the participants to choose the preferred pictures and the reaction time. The reasons were divided into three types: wholeness, detail, and intuition.

3 Results

3.1 Emotion Induced

To examine the effect of emotion induced, we used the t-test to compare the difference in score of the Chinese version of positive emotion scale. There was no significant difference in the score of positive emotion scale between the pretest score and posttest score of neutral emotion. There was significant difference in the score of positive

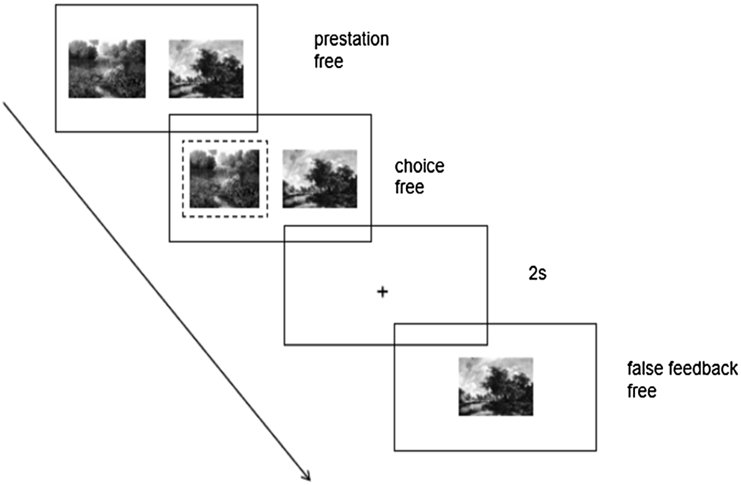


Fig. 1. The process of false feedback in the experiment procedure.

emotion scale between the pretest score and posttest score of positive emotion ($F = -7.12$, $P < 0.0001$, Table 1). The positive emotions of the participants are successfully evoked in positive emotion group.

Table 1. The t-test for score of the Chinese version of positive emotion scale in both groups (positive emotion and neutral emotion).

Group	Pretest score (M ± SD)	Posttest score (M ± SD)	F	P
Positive emotion	2.73 ± 0.73	3.35 ± 0.50	−7.12	0.000***
Neutral emotion	2.77 ± 0.58	2.87 ± 0.66	−1.06	0.30

*: p < 0.05; **: p < 0.01; ***: p < 0.001.

3.2 Choice Blindness

For verifying the existence of choice blindness, we analyzed the detection rate of choice blindness by participants' reports and used SPSS20.0 software to taking statistical analysis. There were the 320 false feedback in total. The detection rate of false feedback was 44.69%. Specifically, it was 56.88% in the positive emotion group, and it was 38.75% in the neutral emotion group. In other words, nearly half of the participants in both groups were not aware of the false feedback. The results showed the choice blindness occurred for Chinese adults really.

3.3 Impacts of Positive Emotion and Picture Similarity on the Choice Blindness

To examine the effect of positive emotion and picture similarity on the choice blindness, we used the chi-square test to compare the difference in number of detections for false feedback pictures. There was significant difference in the detections of false feedback pictures between the two groups ($\chi^2 = 9.70$, $p = 0.008^{**} < 0.01$; Table 2). There was no significant difference between the two groups in high similarity pictures, but there was significant difference between the two groups in low similarity pictures ($\chi^2 = 5.87$, $p = 0.049^{*} < 0.05$; Table 2). The detection rate of false feedback in low similarity pictures was 72% in the positive emotion group, and it was only 49% in the neutral emotion group. Faced with the low similarity pictures, the choice blindness more easily occurred for in the participants of positive emotion group.

Table 2. The chi-square test for number of detecting false feedback pictures with different similarity in both groups.

Group	Picture similarity	0	1	2
Positive emotion	High similarity	13	20	7
	Low similarity	3	16	21
Neutral emotion	High similarity	20	18	2
	Low similarity	11	18	11

Used the chi-square test, there was significant difference between the female and male in both groups ($\chi^2 = 12.98$, $P = 0.04^{*} < 0.05$; Table 3). For further analysis, there was no significant difference between the male in the positive emotion group and the neutral emotion group ($\chi2 = 9.89$, P = 0.14; Table 3). But there was significant difference between the female in the positive emotion group and the neutral emotion group ($\chi^2 = 7.19$, $P = 0.02^{*} < 0.05$; Table 3).

Table 3. The chi-square test for number of detecting false feedback pictures with different gender in both groups

Group	Gender	0	1	2
Positive emotion	Female	5	20	15
	Male	11	16	13
Neutral emotion	Female	14	19	7
	Male	17	17	6

There was significant difference between the female in the positive emotion group and the neutral emotion group when they faced the low similarity pictures ($\chi^2 = 7.19$, $P = 0.02^{*} < 0.05$; Table 4). There existed an interaction effect of emotion, gender and similarity on choice blindness (Fig. 2). Faced the low similarity pictures, the average of female in the positive emotion group was 70%, but the average of female in the neutral emotion group only was 45%.

Table 4. The chi-square test for female number of detecting false feedback pictures with different similarity in both groups

Group	Picture similarity	0	1	2
Positive emotion	High similarity	4	11	5
	Low similarity	1	9	10
Neutral emotion	High similarity	7	11	2
	Low similarity	8	7	5

We recorded the participants' reasons for choosing preferences and divided three part (i.e. wholeness, detail and intuition). There was significant difference of reasons between female in positive emotion group and neutral emotion ($\chi^2 = 8.93$, $P = 0.01^* < 0.05$; Table 5). The female participants in neutral emotion were focus on scenery pictures' wholeness. However, the female participants in positive emotion were focus on scenery pictures' detail.

Table 5. The chi-square test for Reasons of false feedback received by female in both groups

Group	Wholeness	Detail	Intuition
Positive emotion	11	19	10
Neutral emotion	24	9	7

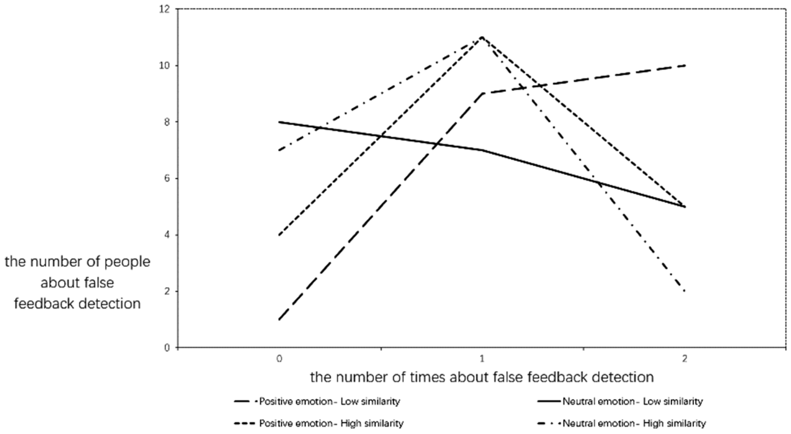


Fig. 2. The interaction effect of emotion, gender and similarity on choice blindness.

4 Discussion

In this study, there was significant difference in the score of positive emotion scale between the pretest score and posttest score of positive emotion. The positive emotion of the participants was successfully evoked in positive emotion group. The participants of positive emotion group were in a positive emotional state. The participants of neutral emotion group maintained emotional stability.

Nearly half of the participants in both groups were not aware of the false feedback. The results showed the choice blindness occurred for Chinese adults really. Computing the result of choice blindness, it is consistent with the previous study in China [9, 10].

In this study, we found that positive emotions also promoted the participants awareness of false feedback. Some researchers studied the effects of emotion on change blindness based on experiments. They found that positive emotions promoted individual awareness of change, while negative emotions hindered individual awareness of change [5]. Change blindness and choice blindness had the similar phenomenon. Whether there are common mechanism of the change blindness and choice blindness has to be studied further. In the studies of emotion, some researchers found positive emotion improved the retrieval speed of individual's attention and enhanced attention [11]. According to motivational dimensional model of affect, some researchers studied the influence of positive emotions on individual cognitive control in different states of approaching motivation intensity by controlling the arousal level of positive emotions. They found that low approach-motivated positive affect enhances cognitive flexibility, whereas high approach-motivated positive affect enhances cognitive stability. This research extends previous work on cognitive breadth to cognitive control which partly reflects the temporality of cognitive processes. Taken together, this line of research suggests that the effects of positive affect on cognitive processes are modulated by its approach motivational intensity [12].

In the classical choice blindness experiment, the researchers used the faces pictures with different similarity as materials and found that the difference in similarity had no significant difference in the detections rate of choice blindness. That is, similarity was not the influencing factor of choice blindness [1]. But in our study, there existed an interaction effect of emotion, gender and similarity on choice blindness. When participants received the feedback of pictures (including veridical and false feedback), their reasons for the preference pictures were focused on the wholeness. No matter how many times the experiment was carried out, it can be clearly seen that the overall description was the main reason. This was consistent with previous studies. Some researchers pointed out in the cross-cultural study of attention that the overall processing mode of Oriental people involved a larger area of attention [13]. In exploring the psychological mechanism of choice blindness, it was found that the overall feature description was more than the specific feature description when the participants gave causal descriptions to the pictures with veridical or false feedback [1]. As for the reasons for false feedback pictures, we found that although both males and females paid more attention to the picture' wholeness, and females focused on the details of the picture was 13.4%. While males paid attention to the details of the picture was 24.5%. This suggested that men pay more attention to detail than women, which may explain why males who were tested in one-week time interval group spent significantly more time than females. Therefore, male participants in one-week time interval group were more likely to detect false feedback.

Future work should increase the number of participants to enhance the sample representativeness. The selection of experimental materials also needs to be more careful and considerable.

5 Conclusions

We studied the impact factors of positive motion and picture similarity in choice blindness based on visual cognitive tests. The results verified the existence of choice blindness in both positive emotion group and neutral emotion group. It was found that significant difference existed in the perception of false feedback between the positive emotion group and neutral emotion group. The female participants in the positive emotion group was more easily perceived the false feedback of low similarity scenery pictures than those female participants in the neutral emotion. The positive emotion had an effect on the choice blindness of the female when they faced the low similarity scenery pictures.

References

1. Johansson, P., Hall, L., Sikstrm, S., Olsson, A.: Failure to detect mismatches between and outcome in a simple decision task. Science **310**(5745), 116–119 (2005)
2. Hall, L., Johansson, P., Trning, B., Sikstrom, S., Deutgen, T.: Magic at the market place: choice blindness for the taste of jam and the smell of tea. Cognition **117**(1), 54–61 (2010)
3. Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., Johansson, P.: How the polls can be both spot on and dead wrong: using choice blindness to shift political attitudes and voter intentions. Public Libr. Sci. **8**(4), 60554–60562 (2013)
4. Masuda, S., Hoshi, S., Watanabe, S.: Choice blindness in the attractiveness of paintings. Ser. Adv. Study Logic Sensib. **4**(1), 81–91 (2010)
5. Pessoa, L., Kastner, S., Ungerleider, L.G.: Attentional control of the processing of neutral and emotional stimuli. Cogn. Brain Res. **15**(1), 31–45 (2002)
6. Zhang, H., Xu, F.M., Xu, M.B.: Choice blindness: did you really know what you have chosen? Adv. Psychol. Sci. **22**(8), 1312–1318 (2014). (in Chinese)
7. Zen, K., Duan, J.Y., Tian, X.M.: The impact of cognitive styles and sensory modalities to choice blindness. Chin. J. Ergon. **22**(4), 5–9 (2016). (in Chinese)
8. Ding, S.W.: Choice blindness: how stable is your choice blindness? Guide Sci. Educ. **3**, 177–178 (2017). (in Chinese)
9. Zhang, H.: The psychological mechanism of choice blindness: based on memory representation perspective. Central China Normal University, Wu Han (2017). (in Chinese)
10. Li, N.: The effect of the presence of others on the choice blindness. Sichuan Normal University, Cheng Du (2017). (in Chinese)
11. Zhang, T.: Theoretical models for the influence of positive emotions on individual attention. Sci. Educ. Article Collects **10**, 137–139 (2016). (in Chinese)
12. Wang, Z.H., Liu, Y., Jiang, C.H.: The effect of low versus high approach-motivated positive affect on cognitive control. Acta Psychologica Sinica **45**(5), 546–555 (2013). (in Chinese)
13. Liu, S.Q., Wang, H.L., Peng, K.P.: Cross-cultural research on attention and its implications. Advances in psychological science **21**(1), 37–47 (2013). (in Chinese)

Color Ergonomics Research in Harsh Environment Under Three Task Modes

Zhiyang Jiang[1,2](✉) and Wenjun Hou[1,2]

[1] School of Digital Media and Design Art,
Beijing University of Posts and Telecommunications, Beijing, China
janejiang595@gmail.com
[2] Beijing Key Laboratory of Network and Network Culture,
Beijing University of Posts and Telecommunications, Beijing, China

Abstract. To investigate how text color can influence the information recognition and cognitive processing of on-board operators and therefore improve operation efficiency, we conducted color ergonomics research on harsh environment under three typical mission scenarios. First, we chose 25 colors based on the color loop of the Ostwald Color System, and used the Simple Reaction Time (SRT) Paradigm to explore the recognition performance of colored blocks on black backgrounds. Then, based on the results of the first experiment, we constructed two types of environment conditions (general environment, harsh environment), conducted a time-stress experiment on three typical task modes (search, calculation, monitor), and analyzed according visual performances of five colors (red, yellow, green, blue and white). The results show that yellow and white are the more suitable colors under the harsh airborne environment, while blue has the worst performance of all colors. As for effect of task scenarios, the impact caused by different colors on efficiency is greater on the recognition tasks compared with the cognitive tasks; As for effect of environment, harsh environment can aggravate the difference of colors, the good (yellow and white) become better and bad (blue and green) become worse than in the general environment.

Keywords: Color ergonomics · Harsh environment · Task modes

1 Introduction

With the significant development of military technology in the area of informatization and intelligence, the airborne mission system has increased its information volume in order to adapt to the increasingly complex combat environment. Reasonable color design of the airborne mission system helps to reduce the workload of the operator, improve the efficiency of recognition and cognitive processing, and ultimately serve the operational effectiveness of the system. Closed airborne cabin is an important workplace and operational environment in high-tech warfare, and the harsh airborne environment can aggravate the physiological and

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 176–186, 2019.
https://doi.org/10.1007/978-3-030-22507-0_14

psychological load of the worker due to the closed space, loud noise, physical vibration, and complex optical environment conditions.

There have been comprehensive studies on the effects of optical environment on color visual performance [4,6], therefore this study focused on the issue of vibration and noise. Vibration environment has a significant impact on human visual performance. Wang [5] discovered the vibration frequency and amplitude has a significant impact on the visual task completion time and the difficulty rating, and the direction of vibration has a significant impact on the visual task completion success rate. The study by Mallick [3] points out that in the case of continuous vibrating, choosing the correct color combination of the text and the background can solve the problem of difficulties of data entry task on laptop computers.

Hearing and vision have a clear interaction compensation effect involving the central nervous system. Gai [1] found that noises can significantly affect the subject's scoring on visual retention, visual response time, curve matching, target tracking, and stereoscopic vision. Liu [2] conducted a visual search task performance study in a speech noise environment.

Most of the researches directly investigated visual ergonomics, and rarely explored the aspect of text color. Moreover, most of them focused on the influence on visual performance of one single environmental variable, and the interaction influence between environmental variables remains unclear. This paper conducted color ergonomics research on harsh environment condition under three task modes, which can help pilot improve information recognition speed and cognitive processing speed, as well as improve their operation efficiency.

2 Experiment 1 - Colored Block Experiment

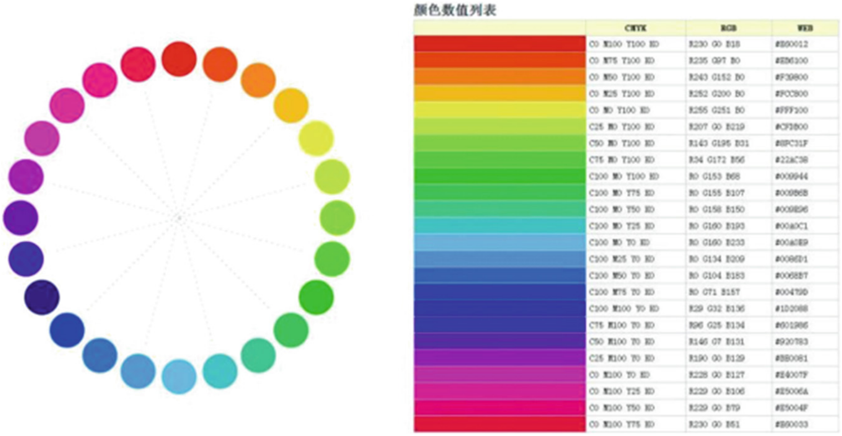
2.1 Experiment Subject

We have 15 male participants as our subject for the colored-block experiment. Their age is between 21–23, average age 21.8 ± 0.6. They are healthy, and have normal vision or corrected-to-normal vision, normal color vision, and normal color perception.

2.2 Experiment Equipment and Material

Computer. We use Lenovo Erazer Y50-70 Laptop: CPU Intel Core i5 4200H, memory 8 GB DDR3, hard drive 256 GB SSD, Graphics Card Nvidia Geforce GTX 860M/Intel GMA 4600, Screen 15.6″ LCD, resolution 1920×1080, color depth 32 bit, refresh rate 60 Hz.

Material. Twenty-five colors were chosen from Ostwald Color System (Fig. 1) together with white. Their index number and their RGB values are 0-(230,0,18), 1-(235,97,0), 2-(243,152,0), 3-(252,200,0), 4-(255,251,0), 5-(207,219,0), 6-(143,195,31), 7-(34,172,56), 8-(0,153,68), 9-(0,155,107), 10-(0,158,150), 11-(0,160,193), 12-(0,160,233), 13-(0,134,209), 14-(0,104,183), 15-(0,71,157), 16-(29,32,136), 17-(96,25,134), 18-(146,7,131), 19-(190,0,129), 20-(228,0,127), 21-(229,0,106), 22-(229,0,79), 23-(230,0,51), 24-(255,255,255). Colored block's size is 2.5 cm × 2.5 cm. Experiment interface's background color is black, its RGB value (0,0,0).



颜色数值列表

	CMYK	RGB	WEB
🟥	C0 M100 Y100 K0	R230 G0 B18	#E60012
🟧	C0 M75 Y100 K0	R235 G97 B0	#EB6100
🟧	C0 M50 Y100 K0	R243 G152 B0	#F39800
🟨	C0 M25 Y100 K0	R252 G200 B0	#FCC800
🟨	C0 M0 Y100 K0	R255 G251 B0	#FFF100
🟩	C25 M0 Y100 K0	R207 G0 B219	#CFDB00
🟩	C50 M0 Y100 K0	R143 G195 B31	#8FC31F
🟩	C75 M0 Y100 K0	R34 G172 B56	#22AC38
🟩	C100 M0 Y100 K0	R0 G153 B68	#009944
🟩	C100 M0 Y75 K0	R0 G155 B107	#009B6B
🟩	C100 M0 Y50 K0	R0 G158 B150	#009E96
🟦	C100 M0 Y25 K0	R0 G160 B193	#00A0C1
🟦	C100 M0 Y0 K0	R0 G160 B233	#00A0E9
🟦	C100 M25 Y0 K0	R0 G134 B209	#0086D1
🟦	C100 M50 Y0 K0	R0 G104 B183	#0068B7
🟦	C100 M75 Y0 K0	R0 G71 B157	#00479D
🟦	C100 M100 Y0 K0	R29 G32 B136	#1D2088
🟪	C75 M100 Y0 K0	R96 G25 B134	#601986
🟪	C50 M100 Y0 K0	R146 G7 B131	#920783
🟪	C25 M100 Y0 K0	R190 G0 B129	#BE0081
🟪	C0 M100 Y0 K0	R228 G0 B127	#E4007F
🟥	C0 M100 Y25 K0	R229 G0 B106	#E5006A
🟥	C0 M100 Y50 K0	R229 G0 B79	#E5004F
🟥	C0 M100 Y75 K0	R230 G0 B51	#E60033

Fig. 1. Ostwald Color System and According Parameters (Color figure online)

2.3 Experiment Design

Experiment program is coded with C++ and Qt framework version 5.11. Before experiment starts, the subjects are sufficiently practiced in order to avoid the practice effect. During the experiments, the indoor light environment stays unchanged, subjects are seated on the chair 60 cm away from the display, both eyes keep parallel with the screen. The subject uses standard computer hardware such as standard mouse and standard keyboard to interact with computer.

Fixed-size colored block was randomly displayed in 25 colors at the center of the screen. Each subject was asked to look at the center of the screen, and press the space bar as soon as they noticed and recognized the color. The colored block will not disappear until the subject has reacted. The interval of the appearance of each colored block was random, ranging between 1–3 s. Each color was repeated 4 times, and each subject performed a total of 100 reactions. Their reaction time were recorded. During the experiment, the subject must be focused, and a reaction is only recorded if it is made after the colored block appeared, meaning pressing the space bar in advance is prohibited and also useless.

2.4 Experiment Results

We analyze the results from the experiment of the reaction time. The abnormal values are omitted from the analysis, then the mean value of each color are calculated and shown in Fig. 2.

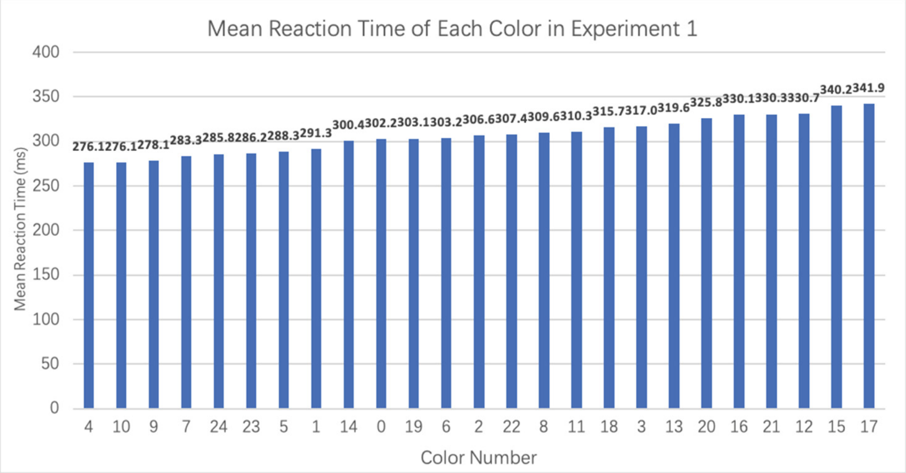


Fig. 2. Mean reaction time of each color in experiment 1 (Color figure online)

From Table 1, analysis of variance showed that there was a significant difference in reaction time of the 25 colors ($F = 1.84, P < 0.05$). Least Significant Difference (LSD) of the 25 colors showed that, the blue-violet color series (No. 15, 16, 17) have performance significantly slower than the yellow color series. Also, red, yellow, cyan and white has the tendency of faster than other colors. The difference in other colors' performance is not significant.

Table 1. ANOVA: the effect of colors on completion time

TIME					
	Sum of squares	df	Mean square	F	Sig.
Between groups	381341.036	24	15889.210	1.846	.008
Within groups	8392924.800	975	8608.128		
Total	8774265.836	999			

Based on these results, also considering the integrity of color series, five colors were chosen as the color for the interface in the following experiment 2: red (230,0,18), yellow (252,251,0), green (0,155,107), blue (0,134,209) and white (255,255,255).

We also studied the count of the color with the shortest reaction time (Fig. 3) and longest reaction time (Fig. 4) within each experiment group (of 25 experiments). In the shortest reaction time count, the top color is yellow (5 times), accounting for 12% of the total, followed by white. In the longest reaction time count, blue color series (dark blue, light blue, sky blue) are the most (12 times), accounting for 30% of the total.

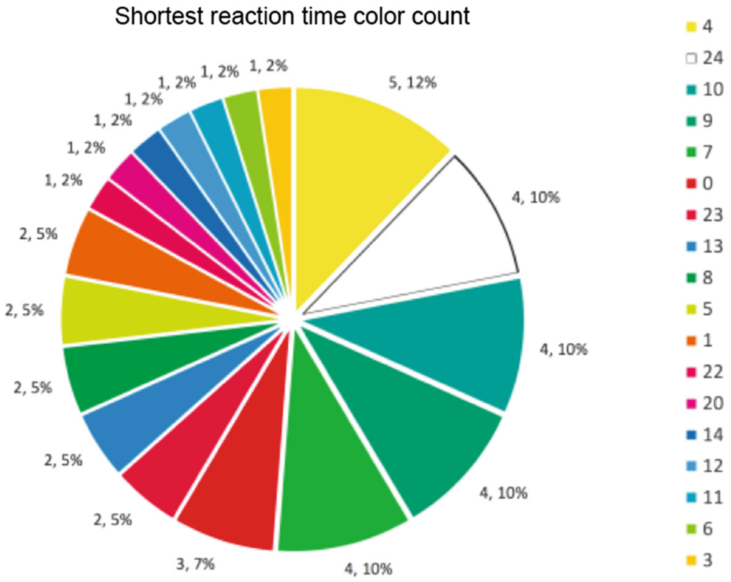


Fig. 3. Count of shortest reaction time for each color (Color figure online)

3 Experiment 2 - Color Ergonomics Experiment

3.1 Experiment Subject

We have 10 male participants as our subject for this experiment. Their age is between 21–23, average age 22.3 ± 0.4. They are healthy, and have normal vision or corrected-to-normal vision, normal color vision, and normal color perception.

3.2 Experiment Equipment and Material

Computer. We use Lenovo Erazer Y50-70 Laptop: CPU Intel Core i5 4200H, memory 8 GB DDR3, hard drive 256 GB Solid State Drive, Graphics Card Nvidia Geforece GTX 860M/Intel GMA 4600, Screen 15.6″ LCD, resolution 1920 × 1080, color depth 32 bit, refresh rate 60 Hz.

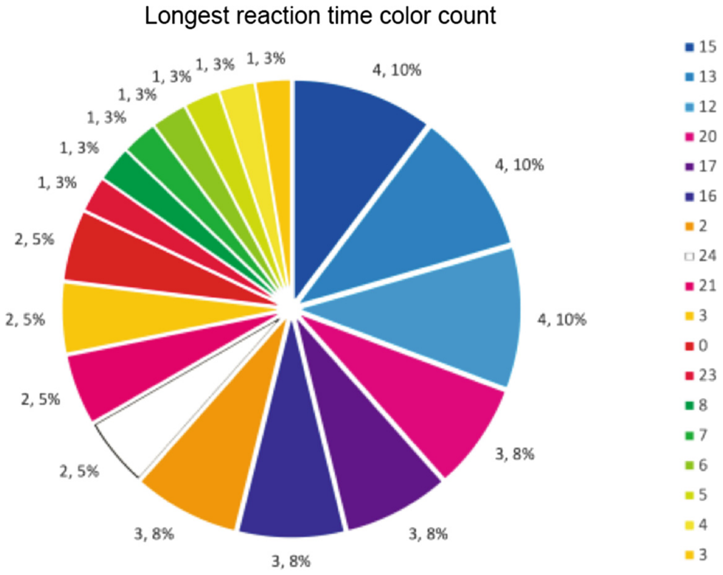


Fig. 4. Count of longest reaction time for each color (Color figure online)

Material. Experiment material used in this experiment are shown in Fig. 5. The three parts on the interface are (from left to right respectively) search task, calculation task and monitor task. We use five colors chosen from Experiment 1 as the interface color: red (230,0,18), yellow (252,251,0), green (0,155,107), blue (0,134,209) and white (255,255,255). The noise is simulated by the software *Soundmasker*.

3.3 Experiment Design

Experiment program is coded with C++ and Qt framework version 5.11. Before experiment starts, the subjects are sufficiently practiced in order to avoid the practice effect. During the experiments, the indoor light environment stays unchanged, subjects are seated on the chair 60 cm away from the display, both eyes keep parallel with the screen. The subject uses standard computer hardware such as standard mouse and standard keyboard to interact with computer.

During the experiment, the subject needs to complete three type of tasks.

- Search task (T1): 25 randomly-chosen different English letters will appear in a 5 × 5 matrix in random order, and one of the letters will be randomly chosen as the target letter. Subject should find and click the target letter in the matrix as soon as possible.
- Calculation task (T2): Two-digit to two-digit add and subtract tasks will be randomly generated and displayed on the interface. The subject need to input the calculated result by keyboard and click “enter” to submit the answer.

- Monitor task (T3): The program will randomly display words from four major categories (animals, plants, food, city names) in a 5 × 3 matrix. We randomly chose one category as the target category. During each task, there will be only zero or one word of the target category. Subjects need to monitor if any word from the target category is shown in the matrix. If there is, the subject needs to click on it; if not, the subject do not need to react.

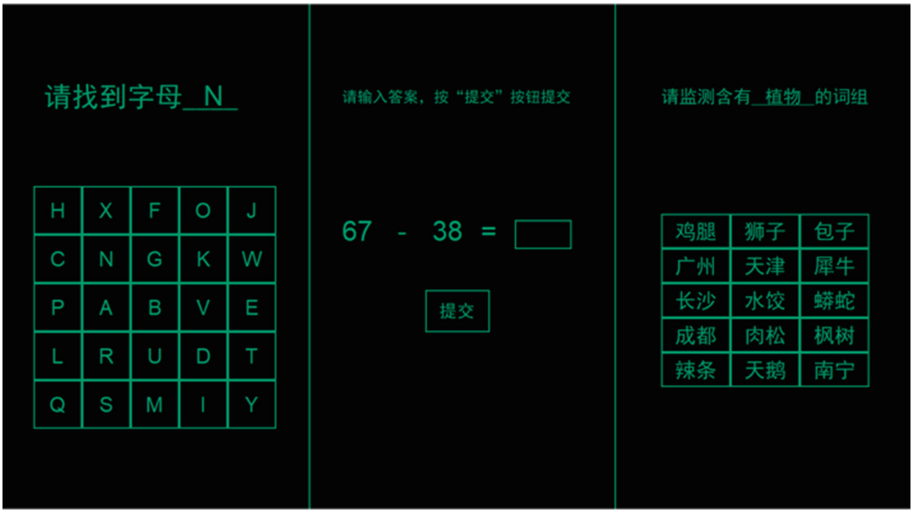
To simulate different environment scenarios, the experiment designed different vibration levels and noise levels. Vibration level includes no vibration and high vibration (frequency 20 Hz, amplitude 2 mm) to the graphic interface. Noise level includes low noise (45 dB white noise) and high noise (80 Hz white noise).

Each subject was asked to complete tasks in two environments.

- General Environment (E1): low noise (45 dB) with no vibration.
- Harsh Environment (E2): loud noise (80 dB) with high vibration (frequency 20 Hz, amplitude 2 mm).

There were five task groups in each environment corresponding to the five colors selected from Experiment 1. There were four task page within each task group, and each task page consisted of a search task, a calculation task, and a monitor task. The operation completion time and success rate of the subject were recorded.

Therefore overall, each subject needs to complete 40 search tasks, 40 calculation tasks and 40 monitor tasks. For monitor tasks, however, the subject need not to react if there are no words from the target category, therefore no reaction time on these tasks. We programmed 10 out of 40 tasks to contain words from the target category. That means, for each subject we have the experiment results for 40 search tasks, 40 calculation tasks and 10 monitor tasks.



请找到字母_N_

H	X	F	O	J
C	N	G	K	W
P	A	B	V	E
L	R	U	D	T
Q	S	M	I	Y

请输入答案，按“提交”按钮提交

67 - 38 = □

请监测含有_植物_的词组

鸡腿	狮子	包子
广州	天津	犀牛
长沙	水饺	蟒蛇
成都	肉松	枫树
辣条	天鹅	南宁

Fig. 5. Graphic interface of experiment 2 (Color figure online)

3.4 Experiment Results

The mean completion time of each task under different environments is shown in Fig. 6. The analysis of variance is shown in Table 2.

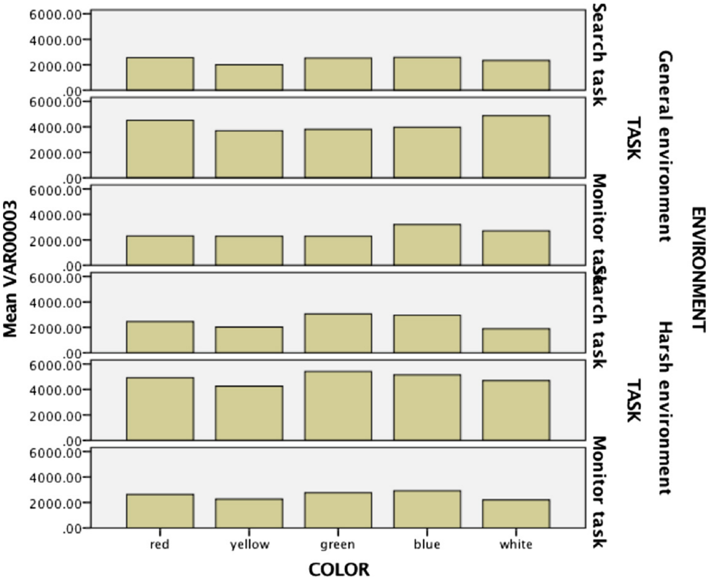


Fig. 6. Mean value of completion time of experiment 2 (Color figure online)

General Environment (Low Noise, No Vibration). In search task, there was significant difference in performance for different colors ($F = 16.122, P < 0.05$). Combined with Least Significant Difference (LSD), the results showed that yellow was significantly better than other colors. In calculation task, there was a trend of significant difference ($F = 1.999, P < 0.1$). In monitor task, there was no significant difference in completion time between colors.

Harsh Environment (Loud Noise, High Vibration). In search task, there was significant difference in performance for different colors ($F = 29.163, P < 0.05$). Combined with histogram and Least Significant Difference (LSD), the completion time of yellow and white was significantly shorter than that of blue and green, and there was no significant difference between red and white. The completion time ranking from fastest to slowest is yellow, white, red, blue, and green. In calculation task, there was a significant difference in performance for different colors ($F = 2.511, P < 0.05$). Blue and green were significantly slower than the other three colors, while the other colors were not significantly different. In monitor task, there was no significant difference in completion time between colors.

Table 2. ANOVA: the effect of colors on completion time, grouped by environment and task

TIME							
Env.	Task		Sum of squares	df	Mean square	F	Sig.
General environment	Search	Between groups	52800285.310	4	13200071.327	16.122	.000
		Within groups	142462010.277	174	818747.185		
		Total	195262295.587	178			
	Calculation	Between groups	16923663.273	4	4230915.818	1.999	.097
		Within groups	370433611.971	175	2116763.497		
		Total	387357275.244	179			
	Monitor	Between groups	8666006.194	4	2166501.548	.755	.562
		Within groups	91842561.050	32	2870080.033		
		Total	100508567.243	36			
Harsh environment	Search	Between groups	94014689.263	4	23503672.316	29.163	.000
		Within groups	132174190.713	164	805940.187		
		Total	226188879.976	168			
	Calculation	Between groups	26437219.242	4	6609304.810	2.511	.043
		Within groups	471140332.215	179	2632068.895		
		Total	497577551.457	183			
	Monitor	Between groups	3286249.244	4	821562.311	.663	.623
		Within groups	35926358.286	29	1238839.941		
		Total	39212607.529	33			

Combined with Table 3 and LSD results, the differences between colors in the general environment and the harsh environment are both significant, and the difference is slightly larger when in the harsh environment. The multivariate analysis of variance (Table 5) also showed that the interaction of the color and environment had a significant effect on the completion time ($F = 3.345, P < 0.05$).

Table 3. ANOVA: The effect of colors on completion time, grouped by environment

TIME						
Environment		Sum of squares	df	Mean square	F	Sig.
General environment	Between groups	56847839.856	4	14211959.964	5.311	.000
	Within groups	1046349945.303	391	2676086.817		
	Total	1103197785.159	395			
Harsh environment	Between groups	103005204.044	4	25751301.011	8.011	.000
	Within groups	1227866758.653	382	3214310.887		
	Total	1330871962.698	168386			

Combined with Table 4 and the LSD results, the difference between colors was significant in Task 1 ($F = 38.244, P < 0.05$), while the difference between Task 2 and Task 3 was not significant. The multivariate analysis of variance (Table 5) also showed that the interaction of color and task type had a significant effect on the completion time ($F = 2.866, P < 0.05$).

Table 4. ANOVA: The effect of colors on completion time, grouped by task

TIME						
Task		Sum of squares	df	Mean square	F	Sig.
Search	Between groups	130567194.003	4	32641798.501	38.244	.000
	Within groups	292755724.583	343	853515.232		
	Total	423322918.586	347			
Calculation	Between groups	14951827.944	4	3737956.986	1.502	.201
	Within groups	893260937.473	359	2488192.026		
	Total	908212765.418	363			
Monitor	Between groups	6514559.549	4	1628639.887	.807	.525
	Within groups	133225481.550	66	2018567.902		
	Total	139740041.099	70			

Table 5. The multi-factor variance analysis of the influence of color, environment and task on completion time

Tests of between-subjects effects Dependent variable: TIME					
Source	Type III sum of squares	df	Mean square	F	Sig.
Corrected model	1213737412.753[a]	29	41853014.233	25.334	.000
Intercept	5616758030.160	1	5616758030.160	3399.912	.000
COLOR	55030226.777	4	13757556.694	8.328	.000
ENVIRONMENT	5834307.184	1	5834307.184	3.532	.061
TASK	976390456.223	2	488195228.112	295.512	.000
COLOR * ENVIRONMENT	22103333.779	4	5525833.445	3.345	.010
COLOR * TASK	37881196.638	8	4735149.580	2.866	.004
ENVIRONMENT * TASK	8100586.592	2	4050293.296	2.452	.087
COLOR * ENVIRONMENT * TASK	6252711.306	8	781588.913	.473	.876
Error	1243979064.511	753	1652030.630		
Total	12746999853.000	783			
Corrected total	2457716477.264	282			

[a] R Squared = .494 (Adjusted R Squared = .474)

4 Conclusion

Color Performance. For different tasks and different environments, color performance has a clear trend. As for the speed of recognition process and cognitive process, yellow and white performs considerably better than blue and green, and red is of medium performance. This indicates that. Both experiment 1 and 2 showed that the blue has the lowest performance in recognition and cognitive.

Impact of Color in Different Environments. The selection of color has less effect in the general environment than in harsh environment. When the environment worsened, people are more sensitive to color, and the differences between the performances of colors are magnified. This shows that in severe environments, the choice and application of color needs to be made with more caution.

Impact of Color on Different Task Modes. Color selection has a significant impact on the search task, but not on calculation and monitor tasks. A search task is a process of identification and matching which involves little processing, while the calculation and monitor tasks have longer information processing time, thus the effect of color is weakened. Therefore, when applying of principles of color selection, it is best to focus on recognition rather than cognition. Color selection should be designed with great effort for recognition tasks.

References

1. Gai, Z., Cui, B., She, X., Chen, X., An, G., Ma, Q.: Vision-related work efficiency of operators in environmental noise based on NES-C4. J. Prev. Med. Chin. People's Lib. Army **9**, 002 (2017)
2. Liu, C., Wu, X., Yu, R.: Research on visual search performance in speech noise environment. Ind. Eng. Manag. **19**(5), 134–139 (2014)
3. Mallick, Z.: Investigating data entry task performance on a laptop under the impact of vibration: the effect of color. Int. J. Occup. Saf. Ergon. **13**(3), 291–303 (2007)
4. Wang, W., Ge, L., Li, H.: A study on the relative color-identification ability on a CRT display in the twilight condition. Chin. J. Ergon. **6**(4), 14–17 (2000)
5. Wang, W., Sun, Y., Lin, Y.: The effects of display vibration on visual performance. China Illum. Eng. J. **24**(3), 24–29 (2013)
6. Xu, W., Zhu, Z.: Color temperature and target luminance on color coding in a CRT display. Acta Psychologica Sinica **4**, 004 (1989)

Semi-automatic Aggregation of Multiple Models of Visual Attention for Model-Based User Interface Evaluation

Dennis Knoop[1], Bertram Wortelen[2](✉), and Marcus Behrendt[2]

[1] Carl v. Ossietzky University, 26129 Oldenburg, Germany
dennis.knoop@uni-oldenburg.de
[2] OFFIS - Institute for Information Technology, 26121 Oldenburg, Germany
{wortelen,behrendt}@offis.de

Abstract. Predicting the distribution of attention to new user interface designs can provide valuable information during the design process, but accurate predictions are difficult to achieve. Recent studies have shown that accuracy can be increased based on the Diversity Prediction Theorem if multiple, independently developed models for the prediction of attention distribution are aggregated. However, aggregating multiple models is a manual task, that takes a lot of effort because a large number of information sources, which are defined as parts of each model, need to be compared among each other. In this work we test two different clustering approaches for automatically aggregating such models. We show that the clustering quality is not sufficient for fully automatic clustering and present a software-supported solution for a semi-automatic clustering process.

Keywords: Areas of Interest · Modelling attention distribution · Clustering multi-word terms · Clustering shapes · Aggregating visual attention models

1 Introduction

For the design of many technical systems, it can be very valuable to know where people are directing their visual attention. Especially with operators in dynamic and safety-critical systems, such as car drivers or aircraft pilots, the way in which visual attention is distributed is essential for safety. Investigations on how changes to the system (e.g., new assistance systems or cockpit designs) affect attention distribution are therefore commonplace. The knowledge about the attention distribution can be helpful even in non-safety-critical areas, e.g. in the ergonomic design of websites or the optimal placement of advertising.

The examination of the attention distributions is usually done with eye trackers. This is the typical way to get reliable results. However, eye tracking studies

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 187–199, 2019.
https://doi.org/10.1007/978-3-030-22507-0_15

have the disadvantage that they are complex, time-consuming and not always easy to perform. Particularly in safety-critical areas, new systems are difficult to test in the real environment, thus they have to be studied in secure test fields or simulation environments. Predicting the distribution of attention is an alternative solution. There is a large body of visual attention models in the literature that focus on very different aspects of attention [1]. Among them is the SEEV (Salience, Effort, Expectancy, Value) model [2], which is relatively easy to use. It was used, among other things, to predict the distribution of attention among motorists [3,4], pilots [2], seafarers [5], and scrub nurses [6]. Even though using the SEEV model is easy, it can still be difficult to make accurate predictions with it. Information sources (or Areas of Interest) must be defined to apply the SEEV model. Parameters are determined for each information source that describe the influence of the factors *Salience*, *Effort*, *Expectancy*, and *Value*. It has been shown that accuracy of predictions depends strongly on the subjective evaluations of the modeler, which are subject to very high variability and thus also to a high expected error [7].

However, according to the Diversity Prediction Theorem formulated by Page [8], one can make a pretty accurate prediction, even with large individual errors, when a larger number of individual assessments are aggregated. The prerequisite for this is that the individual assessments are independent of each other and are normally distributed around the actual value. In such a case, aggregated prediction of a group of individuals is significantly better than that of one individual. The influence of misjudgements of a single analyst is thereby minimized. This is often referred to as the "Wisdom of the Crowd" effect.

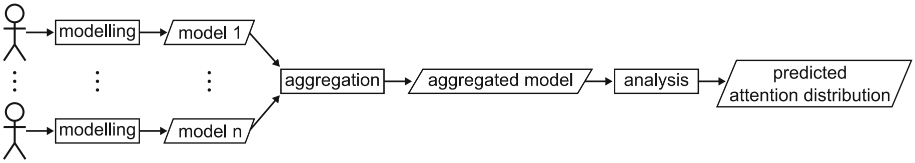


Fig. 1. Modelling process for the multi-model approach. Multiple models are independently created and aggregated. The aggregated model is analyzed to predict attention distribution.

Such an effect was also observed in the prediction of human attention distribution [7]. In a multi-model approach, several people were asked to individually model the attention distribution of a driving situation using the SEEV model. The model forecasts were then aggregated. This process is depicted in Fig. 1. This highly increased the prediction accuracy compared to the average individual prediction [7]. For the aggregation, it is necessary to group the information sources defined by the individual modelers into classes. Figure 2 shows an exemplary information source class. The white rectangles are information sources defined by different people, that all belong to the same information source class *"road ahead"*. Grouping these information sources into classes of information

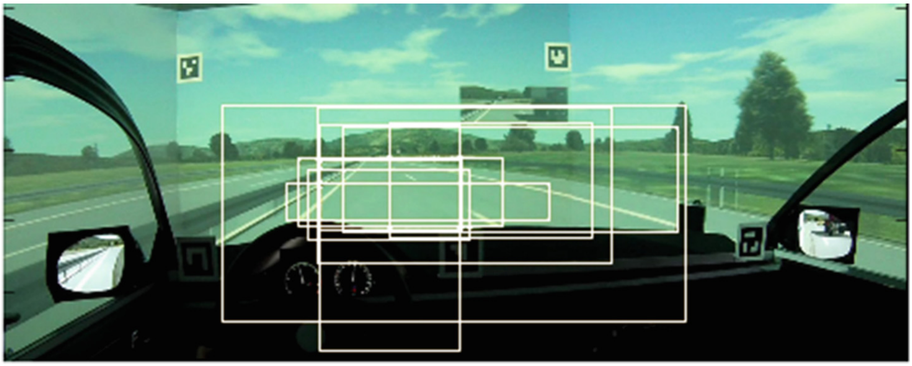


Fig. 2. Information source class *"road ahead"* marked by different modelers for the same situation

sources is a manual process that is problematic for several reasons. One problem is that the process is very time-consuming and doe not scale well. All other aspects of the process depicted in Fig. 1 can either be performed in parallel or are not affected by the number of modelers. In the multi-model approach, several modelers can model at the same time (see Fig. 1). This can considerably reduce the time required for modeling. The number of modelers therefore does not affect the time needed to create the model. Also, the effort for the analysis of the aggregated model is independent of the number of modelers. However, the cost of aggregating the models increases with the number of modelers. The reason for this is the manual grouping of information sources. The effort of all other steps is independent of the number of modelers, or can be automated.

Another problem with grouping is that it is a manual process. It is thus subject to individual errors of the human classifier. However, individual mistakes in modeling are exactly what you want to avoid through the multi-model approach. One solution is to have the classification done by several people and to indicate the quality of the results using a concordance measure such as Fleiss' Kappa [9]. A disadvantage of this solution, however, is that the effort increases even more, because multiple people have to do the same manual classification process.

The problems mentioned above might be one of the reasons, why such a reliable and traceable model-based attention prediction process is rarely adopted for user interface evaluation. The objective of this work is to make the multi-model approach more easily applicable by reducing the effort of manually grouping information sources. For this we present a semi-automated procedure for clustering information sources into classes of information sources. We tested a geometric clustering approach that is based on the region covered by an information source, and a semantic clustering approach that is based on textual labels which were assigned to the information sources by the modelers.

First, we show some general approaches to the cluster analysis and highlight specificities of clustering based on textual descriptions. Then we introduce our clustering algorithm. We evaluate the algorithm using manually clustered reference data in Sect. 3. Next, we discuss what a reasonable practical application of the algorithm in a software tool chain can look like.

2 Clustering of Information Sources

General clustering algorithms for the automatic grouping of objects are well known [10]. A central aspect of a clustering algorithm is a measure of the similarity of objects. This measure is defined over the attributes of the objects to be grouped and is called distance function. The objects are grouped so that the distances within a group are as small as possible and those between groups are large. In hierarchical methods, the objects are either hierarchically subdivided into smaller and smaller groups until the desired number of clusters is reached, or they are merged starting from single element groups until the desired number of clusters is reached. In non-hierarchical methods, an initial grouping is optimized until there is no improvement [11].

To define a suitable distance function for information sources, we are pursuing two approaches that use different attributes of information sources:

Geometry The location and form of information sources that are defined by different modelers but are meant to define the same source should be similar.

Semantics In the modeling of attention distribution and also in eye tracking studies, information sources are typically named or provided with identifiers describing what information can be extracted from the source or what the nature of the information source is. The labels of information sources that were defined by different modelers but are meant to define the same source should be semantically similar.

2.1 Geometry-Based Distance Function

In principle, the shape of an information source can be defined arbitrarily complex. Therefore, different morphometric measures can be used for the distance function. In many cases, sources of information can be described sufficiently accurate by very simple forms. In the studies used for evaluation in this work (Table 1), sources of information are only drawn as rectangles using the attributes x-position, y-position, height (h), and width (w). Since these attributes are cardinal scaled and have the same unit, the Euclidean distance is a good candidate for a distance function of two information sources A and B [11]:

$$d_g(A,B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (w_A - w_B)^2 + (h_A - h_B)^2} \qquad (1)$$

This approach is quite simple and works well for many information source classes. However, as Fig. 2 demonstrates, sometimes the areas of information sources of the same class can differ strongly.

2.2 Semantic-Based Distance Function

To measure the semantic distance of the labels of information sources we used the lexical-semantic nets *WordNet* [12] for English labels and *GermaNet* [13,14] for German labels. GermaNet was built following the example of WordNet and is compatible to it. These nets provide information about the semantic relatedness of single words (nouns, verbs and adjectives) [15]. We used them to calculate a distance between the information source labels. In most of the cases the labels consist of multiple words. Therefor we use a multi word expression distance measure proposed by Huang and Sheng [16].

WordNet is a graph-like structure. Its nodes are sets of synonym words called synsets e.g. the words “foyer” and “lobby” both mean a large entrance or reception room or area. Therefore, they are part of the same synset. Those single words contained in the synsets are called lexical units. Furthermore, each synset has a textual description of its meaning. The edges of the WordNet graph are constructed by ten different semantic relations that connect either two complete synsets (conceptual relations) or two specific lexical units (lexical relations). The four lexical relations of WordNet and GermaNet are synonymy, antonymy, pertonymy and participle. Pertonymy connects adjectives with nouns from which they were derived and the participle relation does the same for adjectives and verbs. Besides these there are the six conceptual relations hyperonymy, hyponymy, meronymy, holonymy, causation and association. Hyperonymy descripes an “is-a” relation between two concepts, e.g. a carrot is a vegetable and the hyponymy describes the opposite relation. Accordingly meronymy and holonymy are also opposites where meronymy describes a part - whole relation, e.g a branch is part of a tree. Inversely the holonymy relation means for example that a tree consists of branches. The causation relation connects verbs with adjectives that express the result of the verb, e.g. the words “(to) close” and “closed” are related because a door is closed after you close it.

In WordNet there are nine measures of semantic relatedness between two synsets. GermaNet implements all of these except for the measure *vector*. These can be distinguished into six similarity measures and two relatedness measures, whereby the similarity measures use only the hyponymy relation and the relatedness measures make use of all relations. The six similarity measures are called *res*, *lin*, *jcn*, *ich*, *wup* and *path*. The *path* measure for example defines the similarity of two synsets as the length of the shortest path along the hyponymy edges between them. Another example is the measure *wup* that uses the least called subsumer (LCS) to calculate the similarity of two synsets. The LCS is the most specific synset that is an ancestor of the two synsets whereby the specificity is measured using the information content of the common ancestor nodes. The measure *wup* defines the semantic similarity of two synsets A and B as follows [17]:

$$s_{\mathrm{wup}}(A,B) = \frac{2 \cdot \mathrm{depth}(\mathrm{LCS}(A,B))}{\mathrm{dist}_{\mathrm{LCS}}(A,B) + \mathrm{dist}_{\mathrm{LCS}}(B,A) + 2 \cdot \mathrm{depth}(\mathrm{LCS}(A,B))} \tag{2}$$

Here, LCS(A, B) denotes the LCS of A and B, depth() means the distance from the root of the hierarchy to a given node and $\text{dist}_{\text{LCS}}(X, Y)$ means the distance from a node X to the LCS of X and Y. Because the similarity measures use only the hyponymy relation they can only be used to measure the distance between synsets of the same word category and not between synsets of different word categories, e.g. a noun synset and a verb synset. The measures *res*, *lin* and *jcn* also use the LCS to compute the similarity of synsets and *ich* is a path based measure like *path* mentioned above.

Besides these, *hso* and *lesk* are the two relatedness measures implemented for WordNet and GermaNet. The *hso* measure uses all relations to search for a path between two synsets whereby the length of the path should not exceed a certain threshold and that changes its direction as rarely as possible. The second relatedness measure analyses the descriptions of the synsets that shall be compared, searches for accordances and calculates the relatedness based on the number of these accordances.

To calculate semantic similarities between multi word expressions we use a measure proposed by Huang and Sheng [16] that uses the similarity measure *wup*. Basically it calculates the pairwise similarity between all words of two expressions using the *wup* measure. If the similarity of two words can not be calculated, i.e. the words don't belong to the same word category an edit distance like the Levenshtein distance is used instead. Let e_1 and e_2 be two multi word expressions. For each word w in e_1 the measure by Huang and Sheng then searches for the highest similarity of w with a word out of e_2. All similarities that are found in this way are summed up and the sum is called *CostSub*. It is also counted how often the highest found distance is zero (*Skip*). After all words of e_1 have been compared to e_2 the same process is repeated vice versa. The similarity between e_1 and e_2 is then calculated as follows using also the total number of words in both expressions (*Total*) and two weight parameters W_{Skip} and W_{Sub} [16]:

$$d_s(e_1, e_2) = 1 - W_{\text{Skip}} * \frac{\text{Skip}(e_1, e_2)}{\text{Total}(e_1, e_2)} - W_{\text{Sub}} * \frac{\text{CostSub}(e_1, e_2)}{\text{Total}(e_1, e_2) - \text{Skip}(e_1, e_2)} \quad (3)$$

We chose $W_{\text{Skip}} = 1$ and $W_{\text{Sub}} = 2.5$ as proposed by Huang and Sheng [16]. The range of values for the calculated similarity is $0 \leq d_s \leq 1$ in which a value of zero means that the expressions e_1 and e_2 have no similarity and a value of one means that the expressions are identical. This measure allows to compare two information sources A and B based on their labels: $d_s(\text{label}(A), \text{label}(B))$.

2.3 Combined Distance Measure

As a third approach we combined the geometric distance measure d_g and the semantic distance measure d_s. The combined distance measure d_c is calculated as the root mean square of d_g and d_s.

Table 1. Overview of the evaluation data sets.

Data set	Models	Situations	Information sources (IS)	IS/situation	IS classes	Rater	Concordance (Fleiss' κ)	Reference	Language[a]
DS 1	19	3	304	5.33	16	3	0.90	[7]	Ger
DS 2	9	3	200	7.41	16	3	0.88	[7]	Ger
DS 3	20	3	374	6.23	37	3	0.81	[18]	Ger
DS 4	40	3	781	6.51	37	3	0.84	[18]	Ger
DS 5	19	3	353	6.19	37	3	0.78	(b)	Ger
DS 6	8	5	378	9.45	21	3	0.89	(c)	Eng
DS 7	6	3	141	7.83	15	3	0.88	[19]	Ger

[a] ger = German, eng = English
[b] Data not published
[c] Data not published. Scenario presented at AutomotiveUI [20]

3 Evaluation

We investigated whether the geometry-based or the semantic-based measure yields better results. We did this based on data from previous studies with manually clustered information sources.

3.1 Data Sets

To evaluate the proposed clustering methods, seven manually grouped data sets were used as test data (see Table 1). The data comes from studies with 6 to 40 modelers who have modeled the distribution of attention in 3 to 5 different situations. These were all automotive studies focused on analysing attention distribution to different HMIs or in different driving situations, like overtaking, parking, and approaching traffic lights. On average, about 5 to 10 sources of information were defined per situation. These were grouped in each study by 3 raters into 13 to 37 classes of information sources. The inter-rater agreement was measured using Fleiss' κ and was always between 0.78 and 0.9. Participants in most studies were Germans and therefore labeled the information sources in German, except for dataset 6, which involved English participants. For the semantic clustering of dataset 6 we used the English WordNet library [12]. For the other data sets we used the German counterpart GermaNet [13,14].

3.2 Results

Our main research interest was to analyze whether a geometry-based distance function or a semantic-based distance function yields better results. For the data sets listed in Table 1 we evaluated three different distance measures: (1) a geometric distance measure based on information source shapes, (2) a semantic measure based on information source labels, and (3) a combined measure. Furthermore we tested it with two different clustering algorithms: (1) a hierarchical cluster algorithm and (2) the K-Medoids algorithm.

We evaluated the clustering quality in reference to the manually clustered data using the adjusted rand index (ARI) [21] as the measure for the clustering

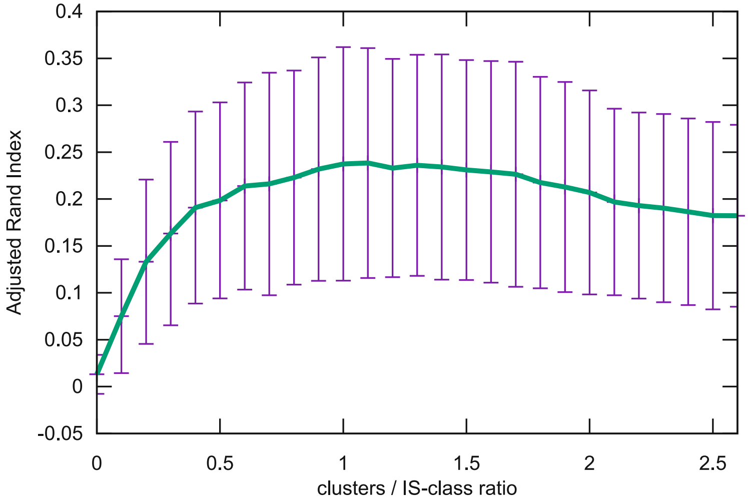


Fig. 3. Mean and standard deviation of adjusted rand index (ARI) over all datasets, clustering algorithms and distance measures.

quality. An ARI of 0 refers to clusters one would get by chance and an ARI of 1 to clusters equivalent to the reference data.

It is in the nature of the data, that the number of clusters is not known in advance, because there is no restriction in what information is marked by participants and especially in what level of detail. Therefore, we first analyzed how the number of clusters created by the algorithms for the different distance measures and datasets affected the clustering quality. Figure 3 shows the mean ARI and standard deviation over all datasets, clustering algorithms and distance measures. The x-axis plots the number of clusters divided by the number of IS classes used in the manual classification. At a cluster/IS-class ratio of 1, as many clusters were generated by the clustering algorithms as IS classes were defined for the manual clustering. In the graph it is easy to see that clustering quality tends to be highest at a ratio of 1. This can also be seen in Fig. 4, which shows the Receiver Operator Characteristic for different cluster numbers (1–100) for all data sets exemplary for the hierarchical algorithm with geometric distance function. At a cluster/IS-class ratio of 1 (marked by the black triangles), both the sensitivity is relatively high and the specificity is high. But it can also be seen that the clustering quality is far from perfect.

Finally, we analyzed our main research question and compared the clustering quality using different distance measures. We used the geometric distance function, the semantic distance function and a combined distance function as described in Sect. 2 to separately cluster all seven datasets. The target number of clusters was always set to the number of IS-classes defined by the raters

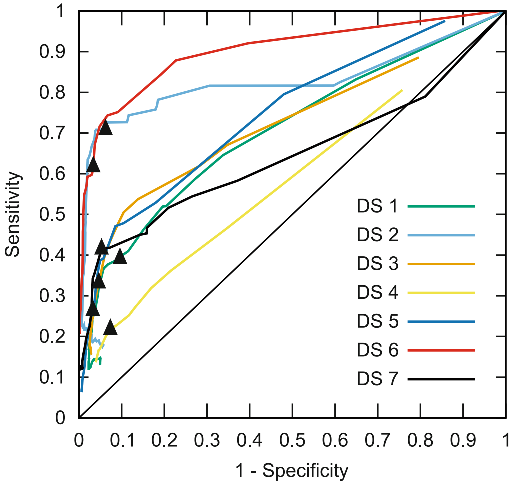
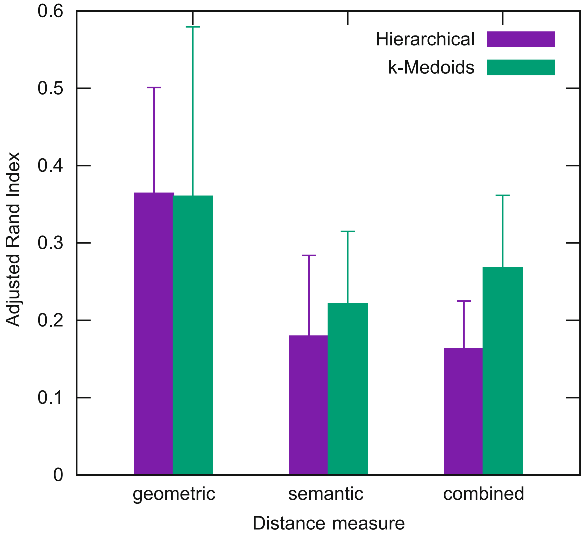


Fig. 4. Exemplary ROC curves for the hierarchical cluster algorithm with geometric distance measure for all seven datasets.

during manual clustering of the respective dataset (cluster/IS-class ratio = 1). The results are shown in Fig. 5.

It can be seen that the geometric distance measure outperforms the semantic distance measure. Combining both measures using an euclidean distance does not result in higher ARI scores than using the geometric distances measure



distance	**Hierarchical**	
measure	mean	(SD)
geometric	0.364	(0.137)
semantic	0.179	(0.104)
combined	0.163	(0.062)

distance	**K-medoids**	
measure	mean	(SD)
geometric	0.360	(0.219)
semantic	0.221	(0.094)
combined	0.268	(0.094)

Fig. 5. Datasets.

alone. The choice of clustering algorithm does not change this general finding. However, the k-medoids algorithm performed better for the semantic and the combined distance measure, but never better than any algorithm relying on the geometric distance measure. However, the geometric distance function does not seem to be equally optimal for all data sets, since the standard deviation of the ARI is significantly higher for the geometric distance function than for the other distance functions. Though it is lower for the hierarchical clustering algorithm compared to the k-medoids algorithm using the geometric distance function. For future applications, it therefore makes sense to use the hierarchical clustering algorithm with geometric distance function.

4 Tool-Support for Semi-automatic Clustering

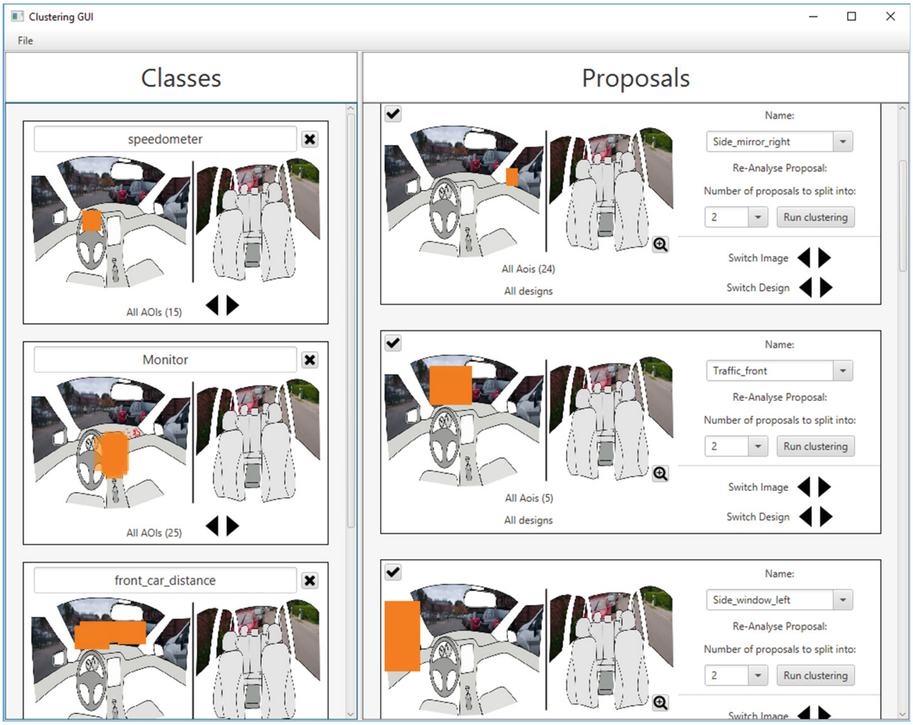
The analysis of clustering quality in the previous section showed, that the clustering quality is far from being perfect. From Fig. 4 it can be seen, that especially the sensitivity is low, meaning that the clustering algorithms often fail to recognize that two information sources should be in the same cluster according to the manually classified information sources. For the application of predictive attention modeling such a high numbers of erroneous classifications is not acceptable. We therefore use the automatic cluster algorithms just as a first step to create initial clusters. We developed a small software application, which visualizes the clusters and allows the user to inspect the clusters and reorganize them. A screenshot of the application is shown in Fig. 6. The tool needs as input the list of information sources and the target number of clusters. It uses the automatic clustering method and displays the clusters as proposals on the right side of the application. The clusters can now be reviewed by inspecting the regions that were marked and also the labels that were provided by the modelers for each information source within a cluster. The user now choose between the three options to

accept the cluster by moving it to the left side of the application that shows the list of IS-classes, which is initially empty.

merge the cluster with an already existing IS-class. This is necessary if the algorithm has created a cluster that the user believes belongs to an already existing IS class. It happened quite often for our data sets. The resulting high number of false negatives is also the reason why the sensitivity of the algorithms is low (see Fig. 4). Merging the clusters is simply done by using drag-and-drop.

split the cluster. If the user believes that the cluster contains elements from more than one IS-class (false positives), then the cluster needs to be divided. The user can do this by rerunning the cluster algorithm, but only on the elements of the current cluster. In the user interface s/he can specify the number of target clusters (typically 2). The original cluster disappears and the new clusters appear in the proposal list on the right side.

This semi-automatic process is far more efficient then the manual clustering process. The most time consuming operation is to split clusters, because it requires



speedometer
All AOIs (15)

Monitor
All AOIs (25)

front_car_distance

Name: Side_mirror_right
Re-Analyse Proposal:
Number of proposals to split into: 2
All Aois (24)
All designs

Name: Traffic_front
Re-Analyse Proposal:
Number of proposals to split into: 2
All Aois (5)
All designs

Name: Side_window_left
Re-Analyse Proposal:
Number of proposals to split into: 2

Fig. 6. Clustering support tool.

to select a target number of clusters and afterwards reviewing the new clusters again and accepting or merging them. Because of this, we suggest to initially select a target number of clusters that is larger then the expected number of IS-classes. The resulting larger number of clusters requires more merging then splitting operations.

5 Discussion

One result of our study is that clustering based on our semantic distance measure is inferior to a geometric distance measure. However, we think that the semantic distance measure has the highest potential for improvement.

For example, WordNet's antonymy relation could be used to identify similar information sources that definitely belong to different IS classes, like *left side mirror* and *right side mirror*.

Another approach could be to use WordNet's meronymy relation to identify different levels of abstraction in the models. One person might create a detailed model by marking the speedometer, the revolution counter, and the fuel gauge. Another person might simply mark the entire dashboard.

The result of this work is a semi-automatic approach for clustering information sources for analyzing models of attention distribution. It turns out that the clustering quality is not sufficient for a fully automatic approach. We therefore created and presented a software for manually revising the automatically created cluster. This approach heavily reduces the clustering effort.

References

1. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 185–207 (2013)
2. Wickens, C.D., Helleberg, J., Goh, J., Xu, X., Horrey, W.J.: Pilot task management: testing an attentional expected value model of visual scanning. Technical report ARL-01-14/NASA-01-7, University of Illinois, Aviation Research Lab, Savoy, IL, November 2001
3. Horrey, W.J., Wickens, C.D., Consalus, K.P.: Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. J. Exp. Psychol.-Appl. **12**(2), 67–78 (2006)
4. Lemonnier, S., Brémond, R., Baccino, T.: Gaze behavior when approaching an intersection: dwell time distribution and comparison with a quantitative prediction. Transp. Res. Part F: Traffic Psychol. Behav. **35**, 60–74 (2015)
5. Feuerstack, S., Wortelen, B.: The human efficiency evaluator: a tool to predict and explore monitoring behaviour. Kognitive Systeme **2017**(1), 11 p. (2017)
6. Koh, R.Y.I., Park, T., Wickens, C.D., Ong, L.T., Chia, S.N.: Differences in attentional strategies by novice and experienced operating theatre scrub nurses. J. Exp. Psychol.: Appl. **17**(3), 233–246 (2011)
7. Feuerstack, S., Wortelen, B.: A model-driven tool for getting insights into car drivers? Monitoring behavior. In: Proceedings of 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 861–868. IEEE (2017)
8. Page, S.E.: The Difference: How the Power of Diversity Creates Better Groups, Schools, and Societies. Princeton University Press, Firms (2008)
9. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378–382 (1971)
10. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (2005)
11. Bortz, J., Schuster, C.: Statistik für Human- und Sozialwissenschaftler, 7th edn. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12770-0
12. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
13. Henrich, V., Hinrichs, E.: GernEdiT - the GermaNet editing tool. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), pp. 2228–2235 (2010)
14. Hamp, B., Feldweg, H.: GermaNet - a lexical-semantic net for German. In: Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications (1997)
15. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet: similarity - measuring the relatedness of concepts. In: Demonstration Papers at HLT-NAACL 2004. Association for Computational Linguistics (2004)
16. Huang, G., Sheng, J.: Measuring similarity between sentence fragments. In: Proceedings of 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (2012)

17. Wu, Z., Martha P.: Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138. ACL (1994)
18. Wortelen, B., Feuerstack, S.: Comparing the input validity of model-based visual attention predictions based on presenting exemplary situations either as videos or static images. In: Proceedings of 15th International Conference on Cognitive Modelling (ICCM 2017) (2017)
19. Feuerstack, S., Wortelen, B., Kettwich, C., Schieben, A.: Theater-system technique and model-based attention prediction for the early automotive HMI design evaluation. In: Green, P., Boll, S., Burnett, G., Gabbard, J., Osswald, S. (eds.) AutomotiveUI 2016: Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 19–22. ACM, New York (2016)
20. Feuerstack, S., Wortelen, B.: Tutorial: how does your HMI design affect the visual attention of the driver. In: Adjunct Proceedings of the 9th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2017), pp. 28–32. ACM (2017)
21. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)

Research on Aesthetics Degree Optimization Model of Product Form

Ming Li, Jie Zhang(✉), and Yiping Hou

East China University of Science and Technology, Shanghai 200237, China
850137117@qq.com

Abstract. A good product design should be a beautiful design. In order to evolve product form based on aesthetic feeling, a comprehensive evolution method of form aesthetic degree was established based on BP neutral network and genetic algorithm. Firstly, the index knowledge of aesthetic degree was constructed and representative aesthetics degree index was selected for target product, thus index dimension were reduced and then validate the rationality of cognition logic through cluster analysis and structural equation model. Secondly, BP neutral network was used to construct aesthetic degree evaluation system of product form, so that it can be built relationship between explicit product form and implicit aesthetic sense. Finally, The aesthetics degree optimization prototype system was build through combine evaluation model with evolution model by BP neutral network and genetic algorithm, respectively. Taking the teapot as a research case, the result showed that TADOPS system could provide effective aids for product.

Keywords: Product form · Aesthetic cognition · Back propagation neutral network · Genetic algorithm · Aesthetic degree optimization

1 Introduction

With the great enrichment of products, consumers' demand for products is deeper and more diversified, and the requirements for aesthetics are getting higher and higher. The user's consumption concept has entered an era where more emphasis is placed on emotional experience. Form is the most direct way to reflect perceptual factors, while aesthetics is the most basic emotional demand of consumers. Therefore, more and more companies pay attention to design and aesthetics to improve the market competitiveness of products.

A good design should first be a beautiful design. What is beauty? Where does aesthetic come from? People generally explore the characteristics and essence of aesthetic from two aspects. On the one hand, it explores the characteristics and essence of aesthetic from the attributes of objective objects. For example, Aristotle in ancient Greece believed that aesthetic depends on volume and order, and aesthetic lies in the unity of various elements. On the other hand, from the philosophical connotation, it is to explore the essence of aesthetic from the human psychology. For example, Plato, the ancient Greek philosopher, believed that aesthetic is a philosophy, and the beauty created by people comes from inner wisdom and love. These ancient aesthetic theorists

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 200–213, 2019.
https://doi.org/10.1007/978-3-030-22507-0_16

showed the signs and principles of aesthetic at a certain level, but the meaning of aesthetic has limitations. Contemporary aesthetic theorists put forward: aesthetic is the unity of objective regularity and social purpose. Aesthetic is objective, and aesthetic feeling is an inner psychological feeling for people. Aesthetic and aesthetic feeling are the products of social history. In recent decades, many scholars have begun to pay attention to the role of aesthetic in emotional design, the influence of aesthetic in the product design process, and how to design products that satisfy consumers' inner emotional appeals.

The more mature theory of consumer aesthetic evaluation research is Kansei Engineering, which was originally proposed by Japanese scholar nagamachi [1], to explore the potential connection between "human" and "object". A large number of scholars have applied sensory engineering to establish product shape design methods and selection methods for different problems. In the evaluation of applied beauty, most of them adopt low-dimensional aesthetic evaluation. Birkhoff who is the founder of computational aesthetics proposed a mathematical model of aesthetic degree in 1933, and the "Aesthetic measure" was expressed as the ratio of order to complexity, i.e. M = O/C [2]. Talia and Noam described the aesthetic characteristics through relevant professionals, and selects 41 aesthetic characteristics indicators, and uses the evaluation value of each index to determine the weight of the indicators, thus completing the comprehensive evaluation [3]. Schenkman and Jonsson evaluated the interface of the web page with a single indicator of "aesthetic" [4]. Furthermore, Some scholars use M = O/C to carry out research on form beauty, like Ngo and teo established a relevant formula for calculating the layout elements in the interface design, which realizes the quantification of the interface layout features [5]. Hsiao and Chou used the fuzzy information entropy method to construct the aesthetic cognitive model of web design which based on the principle of gestalt psychology [6]. At present, the aesthetic study of product form is more focused on aesthetic evaluation, and most of the aesthetic evaluation of product forms are subjective based on expert experience and questionnaires to determine weights. Besides, There are few studies that combine aesthetic evaluation to optimize product morphology. This paper aims at developing of product form intelligent design system based on consumers' aesthetic and market position, which will make the product appearance more suitable for the consumers' individual requirements. We analyzed aesthetic principles of visual perception and cognition, and summarized function knowledge base of aesthetic degree evaluation of product form, then the relationship between aesthetic degree quantization index and aesthetic sense was built, furthermore, we built product form intelligent design prototype system based on biological evolution theory, that provided an effective aided model for the product design and development.

2 Product Form and Aesthetic

The product form is a detailed expression of the designer's aesthetic thoughts, and is also a detailed expression of the functional requirements and aesthetic needs of the designed products. The designer's aesthetic thought, creative value and spiritual connotation should be finally transformed into form. Only through the potential functional meaning and spiritual connotation of the product form can the user perceive and be aware. Therefore, product form design has an irreplaceable position in the industrial design process.

The aesthetic is a combination of attributes such as people, material products, spiritual products, and works of art. It has a variety of characteristics, such as symmetry, proportionality, harmony, vividness, novelty, form suitability, color harmonious, formal integrity and consistency of form, and so on. Aesthetic has the same social attributes as other objects, so it has its own characteristics. It appears and exists in the development of society. It is objectively dependent on things, and is not determined by the image and desire of individual subjective feelings. Aesthetic is judged by the social scale (criteria, thoughts) that human beings have accumulated in various concrete forms in the development of social history. The connotation of beauty is changing, and the emergence of new and more complete things will inevitably lead to the re-evaluation of aesthetic standards by society.

The product form aesthetic brings the sense of beauty to the user and affects the user's reaction, emotion and cognition. It plays an irreplaceable position in the enterprise product strategy and is the key link to obtain the user's favor and the market leader. The product form design must not only satisfy the function of the product, but also convey the needs of spiritual needs, cultural connotations and aesthetic experience. The internal relationship between the product form and the consumer aesthetic experience mechanism, how the aesthetic experience affects the product beauty evaluation, and how the enterprise guides the product design from the aesthetic point of view are all problems that need to be solved. Usually in the product design process, the designer uses the traditional aesthetic tacit knowledge to be subjective, random and fuzzy. Therefore, how to objectively describe the beauty of the product, how to scientifically express the beauty of the product, how to quantify the beauty of the product, and how to optimize the beauty of the product form will be a new exploration in the field of product design.

3 Aesthetic Measures

The formal beauty law is the experience summary and abstract of formal in the process of creating beauty, some of its usage include balance and equilibrium, contrast and blend, change and unity, cadence and rhythm, etc. [7]. Studying and exploring the effects of formal beauty law on human aesthetic perception can guide people to better create beautiful things. In this study, ten indexes were proposed for the aesthetic of product form.

3.1 Degree of Balance

The degree of balance is computed as the difference between center of gravity of components on each side of the X-axis, Y-axis and Z-axis and is given by

$$DB = 1 - \frac{\frac{W_L - W_R}{\max(|W_L|,|W_R|)} + \frac{W_T - W_D}{\max(|W_T|,|W_D|)} + \frac{W_F - W_B}{\max(|W_F|,|W_B|)}}{3} \quad (1)$$

where L, R, T, D, F, and B stand for left, right, top, bottom, face, and back, respectively.

3.2 Degree of Equilibrium

The degree of equilibrium is computed as the difference between weight of components on each side of the horizontal and vertical axis and is given by

$$DE = 1 - \frac{\frac{J_T - J_D}{\max(|J_T|,|J_D|)} + \frac{J_L - J_R}{\max(|J_L|,|J_R|)}}{2} \tag{2}$$

with

$$J_j = S_j D_j, \quad j = T, D, L, R \tag{3}$$

where T, D, L, R, and T stand for top, bottom, left, and right, respectively; S_j is the cross-sectional area on side j; D_j is the distance between the central lines of the object on each side and the whole product.

3.3 Degree of Unity

The degree of unity, by definition, is computed as the relationship between product components and product unity and the compactness of the distribution of product components, and given by

$$DU = \frac{\left|1 - \frac{n-2}{n}\right| + \left|\frac{\max(u_i) - \min(u_i)}{u_{\text{product}}}\right| + \left|\frac{\sum_i^n a_i}{a_{\text{frame}}}\right|}{3} \tag{4}$$

where n is the number of product components; u_i and $u_{product}$ are the volume of object i and product, respectively; a_i and a_{frame} are the areas of object i and cross-section of product.

3.4 Degree of Coordinate

The degree of coordinate is computed as the difference between the physical center of product components and the physical center of product unity and given by

$$DC = 1 - \frac{\left|\frac{2\sum_i^n a_i(X_i - X_c)}{b_{\text{frame}}\sum_i^n a_i}\right| + \left|\frac{2\sum_i^n a_i(Y_i - Y_c)}{h_{\text{frame}}\sum_i^n a_i}\right|}{2} \tag{5}$$

where (X_i, Y_i) and (X_c, Y_c) are the coordinates of object i and product center, respectively; b_{frame} and h_{frame} are the width and height of product, respectively.

3.5 Degree of Deviation

The degree of deviation is computed the deviation of the volume and area of a product from similar products and given by

$$DD = \frac{\left(\sum_{1}^{n} V_i - V_i\right) + \left(\sum_{1}^{n} S_i - S_i\right)}{2} (i \leq N) \tag{6}$$

where N is the number of study samples; Vi and Si are the volume and area of object i.

3.6 Degree of Economy

The degree of economy, by definition, is a measure of how economical the product is and is given by

$$DEY = 1 - \frac{\frac{1}{N_{size}} + \frac{1}{N_{material}} + \frac{1}{N_{object}}}{3} \tag{7}$$

where N_{size}, $N_{material}$ and N_{object} are the number of sizes, materials and components of product, respectively.

3.7 Degree of Homogeneity

The degree of homogeneity, by definition, is a measure of how evenly the objects are distributed among the quadrants and is given by

$$DH = -k \ln \frac{n!}{n_{\mathrm{TL}}!n_{\mathrm{TR}}!n_{\mathrm{DL}}!n_{\mathrm{DR}}!} \tag{8}$$

where k is a constant, known as Boltzmann's constant; n is the number of product components; n_{TL}, n_{TR}, n_{DL}, and n_{DR} are the numbers of objects on the top-left, top-right, down-left, and down-right quadrants, respectively.

3.8 Degree of Symmetry

The degree of symmetry, by definition, is the extent to which the product is symmetrical in three directions: vertical, horizontal, and diagonal and is given by

$$DS = 1 - \frac{|\mathrm{S}_{\mathrm{vertical}}| + |\mathrm{S}_{\mathrm{horizontal}}| + |\mathrm{S}_{\mathrm{radial}}|}{3} \tag{9}$$

where $\mathrm{S}_{\mathrm{vertical}}$, $\mathrm{S}_{\mathrm{horizontal}}$, and $\mathrm{S}_{\mathrm{radial}}$ are, respectively, the vertical, horizontal, and radial symmetries with

$$S_{\text{vertical}} = \frac{1}{8}\left[\frac{|X'_{TL}-X'_{TR}|}{\max(X'_{TL},X'_{TR})} + \frac{|X'_{DL}-X'_{DR}|}{\max(X'_{DL},X'_{DR})} + \frac{|Y'_{TL}-Y'_{TR}|}{\max(Y'_{TL},Y'_{TR})} + \frac{|Y'_{DL}-Y'_{DR}|}{\max(Y'_{DL},Y'_{DR})} + \right.$$
$$\left.\frac{|\theta'_{TL}-\theta'_{TR}|}{\max(\theta'_{TL},\theta'_{TR})} + \frac{|\theta'_{DL}-\theta'_{DR}|}{\max(\theta'_{DL},\theta'_{DR})} + \frac{|R'_{TL}-R'_{TR}|}{\max(R'_{TL},R'_{TR})} + \frac{|R'_{DL}-R'_{DR}|}{\max(R'_{DL},R'_{DR})}\right] \tag{10}$$

$$S_{\text{horizontal}} = \frac{1}{8}\left[\frac{|X'_{TL}-X'_{DL}|}{\max(X'_{TL},X'_{DL})} + \frac{|X'_{TR}-X'_{DR}|}{\max(X'_{TR},X'_{DR})} + \frac{|Y'_{TL}-Y'_{DL}|}{\max(Y'_{TL},Y'_{DL})} + \frac{|Y'_{TR}-Y'_{DR}|}{\max(Y'_{TR},Y'_{DR})} + \right.$$
$$\left.\frac{|\theta'_{TL}-\theta'_{DL}|}{\max(\theta'_{TL},\theta'_{DL})} + \frac{|\theta'_{TR}-\theta'_{DR}|}{\max(\theta'_{TR},\theta'_{DR})} + \frac{|R'_{TL}-R'_{DL}|}{\max(R'_{TL},R'_{DL})} + \frac{|R'_{TR}-R'_{DR}|}{\max(R'_{TR},R'_{DR})}\right] \tag{11}$$

$$S_{\text{radial}} = \frac{1}{8}\left[\frac{|X'_{TL}-X'_{DR}|}{\max(X'_{TL},X'_{DR})} + \frac{|X'_{TR}-X'_{DL}|}{\max(X'_{TR},X'_{DL})} + \frac{|Y'_{TL}-Y'_{DR}|}{\max(Y'_{TL},Y'_{DR})} + \frac{|Y'_{TR}-Y'_{DL}|}{\max(Y'_{TR},Y'_{DL})} + \right.$$
$$\left.\frac{|\theta'_{TL}-\theta'_{DR}|}{\max(\theta'_{TL},\theta'_{DR})} + \frac{|\theta'_{TR}-\theta'_{DL}|}{\max(\theta'_{TR},\theta'_{DL})} + \frac{|R'_{TL}-R'_{DR}|}{\max(R'_{TL},R'_{DR})} + \frac{|R'_{TR}-R'_{DL}|}{\max(R'_{TR},R'_{DL})}\right] \tag{12}$$

X'_j, Y'_j, θ'_j, and R'_j are, respectively, the normalised values of

$$X_j = |X_j - X_c|, j = UL, UR, DL, DR \tag{13}$$

$$Y_j = |Y_j - Y_c| \tag{14}$$

$$\theta_j = \left|\frac{Y_j - Y_c}{X_j - X_c}\right| \tag{15}$$

$$R_j = \sqrt{(X_j - X_c)^2 + (Y_j - Y_c)^2} \tag{16}$$

where TL, TR, DL and DR stand for top-left, top-right, down-left and down-right, respectively; (x_j, y_j) and (x_c, y_c) are the co-ordinates of the centers of product components on quadrant j and the product unity.

3.9 Degree of Proportion

The Degree of proportion, by definition, is the comparative relationship between the dimensions of the proportional shapes and is given by

$$DP = t_j, \min\left(|t_j - t|, j = sq, r2, gr, r3, ds\right) \tag{17}$$

with

$$t = \begin{cases} r, & r \leq 1 \\ \frac{1}{r}, & r > 1 \end{cases}, \quad r = \frac{H}{B} \tag{18}$$

where H and B are the width and height of the product. p_j is the proportion of product j with

$$\{p_{sq}, p_{r2}, p_{gr}, p_{r3}, p_{ds}\} = \left\{\frac{1}{1}, \frac{1}{1.414}, \frac{1}{1.618}, \frac{1}{1.732}, \frac{1}{2}\right\} \tag{19}$$

where sq, r2, gr, r3, and ds stand for square, square root of two, golden rectangle, square root of three, and double square, respectively.

3.10 Degree of Order

Degree of order, by definition, is the extent to which the objects are systematically ordered and is given by

$$D_{\mathrm{s,q}} = 1 - \frac{|R_X| + |R_Y| + |R_{\mathrm{A}}| + |R_{\mathrm{V}}|}{4} \tag{20}$$

with

$$\begin{aligned} R_I = \frac{1}{6}\Bigg[&\frac{I'_{TL}}{\max(I'_{TL}, I'_{UR})} + \frac{I'_{TL}}{\max(I'_{TL}, I'_{DL})} + \frac{I'_{TL}}{\max(I'_{TL}, I'_{DR})} + \frac{I'_{TR}}{\max(I'_{TR}, I'_{DL})} \\ &+ \frac{I'_{TR}}{\max(I'_{TR}, I'_{DR})} + \frac{I'_{DL}}{\max(I'_{DL}, I'_{DR})}\Bigg], I = X, Y \end{aligned} \tag{21}$$

$$\begin{aligned} R_J = \frac{n!}{2 \times (n-2)}\Bigg[&\frac{\min(J'_1, J'_2)}{\max(J'_1, a'_2)} + \frac{\min(J'_2, a'_3)}{\max(a'_2, a'_3)} + \frac{\min(J'_3, J'_4)}{\max(J'_3, J'_4)} + \frac{\min(J'_1, J'_3)}{\max(J'_1, J'_3)} \\ &+ \frac{\min(J'_1, J'_4)}{\max(J'_1, J'_4)} + \frac{\min(J'_2, J'_4)}{\max(J'_2, J'_4)} + \cdots + \frac{\min(J'_{n-1}, J'_n)}{\max(J'_{n-1}, J'_n)}\Bigg], \ J = A, \ V \end{aligned} \tag{22}$$

where An and Vn are the cross-sectional area and volume of product components n on each quadrant; X'_j, Y'_j, A'_n and V'_n are respectively, the normalised values of

$$X_j = \sum_{i}^{n_j} |X_{ij} - X_c| \tag{23}$$

$$Y_j = \sum_{i}^{n_j} |Y_{ij} - Y_c|, j = TL, TR, DL, DR \tag{24}$$

where TL, TR, DL and DR stand for top-left, top-right, down-left and down-right, respectively.

4 Construction the Relationship Between Product Form and Aesthetic Indicators

The representative index of the product form aesthetic evaluation is a necessary step for the locator to recognize the target product form, and is also a key indicator of the product form aesthetic evaluation system. Determining the representative index of aesthetic degree can clarify the aesthetic cognition and emotional appeal of the product form, and provide the basis for the follow-up work of the product form aesthetic evaluation system. In the process of index selection, redundancy will inevitably occur. Therefore, we applied semantic differential method, cluster analysis and structural equation model to eliminate redundancy.

The aesthetics is the user's perception of the product, which expresses its own emotional appeal. When the product appears in front of the user, its structural form will stimulate the human brain vision system and reflect it, forming a user's aesthetic evaluation of the product, and the aesthetic is the user's perceptual perception of the product form, and the process cannot be directly recognized. The product form aesthetic evaluation system can establish a mapping relationship between product form and aesthetic indicators, and provide guidance for designers' product design and modification. Therefore, we tried to explore the black box mechanism between product form and aesthetic indicators by BP neural network.

4.1 Case Study

The purple clay teapot is a unique hand-made clay craft which began in the Zhengde period of the Ming Dynasty in China. It is so popular that artistic and practical combination perfect and culture of zen Buddhism with tea, so we selected teapot as example which been choose from masters and workshops in china. 20 representative teapots (Fig. 1) which remodeled by grayscale process have been scored through interviewed 31 students who have design background.

4.2 Representative Indicator of Aesthetic Degree

In order to establish a product form aesthetic degree Optimization model system, Degree of balance, Degree of symmetry, Degree of proportion, Degree of equilibrium, Degree of unity, Degree of coordinate, Degree of economy, Degree of homogeneity, Degree of order and Degree of deviation were analyzed by cluster analysis K-means, which used K-Means clustering of aesthetic quantization matrices through SPSS software in Table 1. The results shown degree of symmetry, degree of deviation, degree of order and degree of proportion as the representative indexes.

Based on the principle of visual perception and logical judgment, the hypothesis model 1 (Fig. 2) and 2 (Fig. 3) was build to validate aesthetic cognition by structural equation model in AMOS software. The structural equation model's χ^2 and RMSEA value are used as the validation parameters, and χ^2 has the advantage of judging the fit of the structural equation model, the smaller the value, the better the significant differences, and RMSEA is not affected by the sample size and the complexity of the

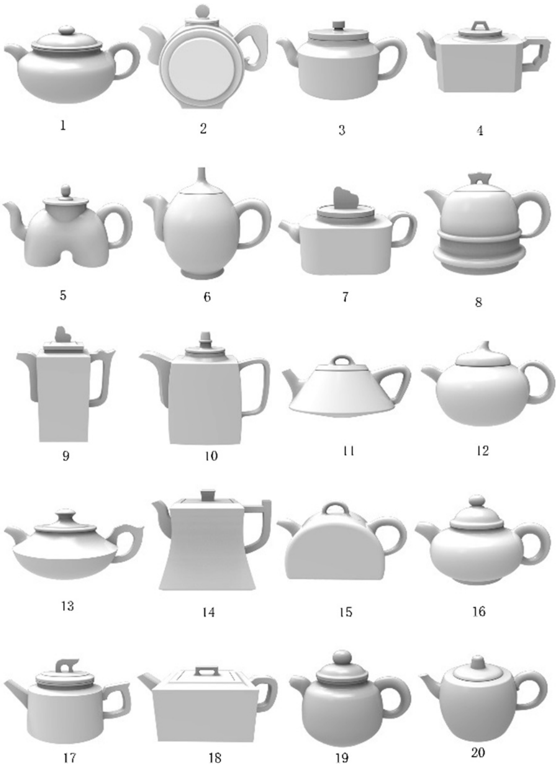


Fig. 1. Twenty representative purple clay teapots

Table 1. Aesthetic degree indexes with cluster analysis

No.	Aesthetic index	Class	Distance
1	Degree of balance	2	0.376
2	Degree of symmetry	2	0.170
3	Degree of proportion	3	0.201
4	Degree of equilibrium	2	0.355
5	Degree of unity	2	0.207
6	Degree of coordinate	2	0.343
7	Degree of economy	3	0.202
8	Degree of homogeneity	2	0.356
9	Degree of order	4	0.000
10	Degree of deviation	1	0.000

model, and the hypothesis model is good in below 0.06. The input variables are aesthetic image through semantic differential method, the hypothesis model 1 is saturated model which χ^2 and RMSEA value is 0 and 0, respectively. Therefore, it is necessary to revise the hypothesis model 1. The hypothesis model 2 has a good result which χ^2 and RMSEA value is 39.857 and 0.06 (below 0.08), respectively, and the degree of freedom is 17, in addition, r1, r2 and e1is error term. The results shown that model of 4 representational indexes is reasonable and reliable.

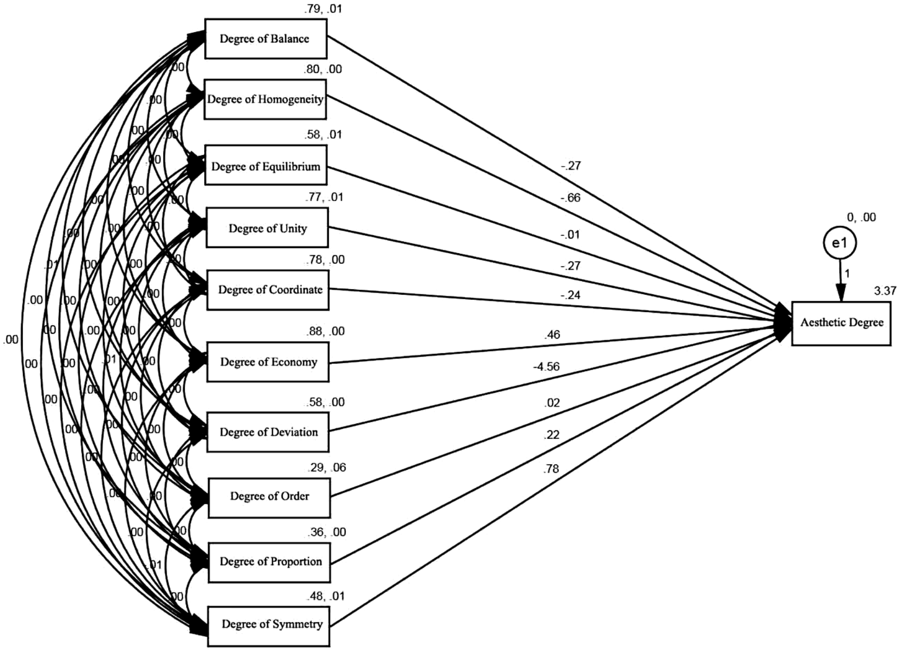


Fig. 2. The hypothesis model 1

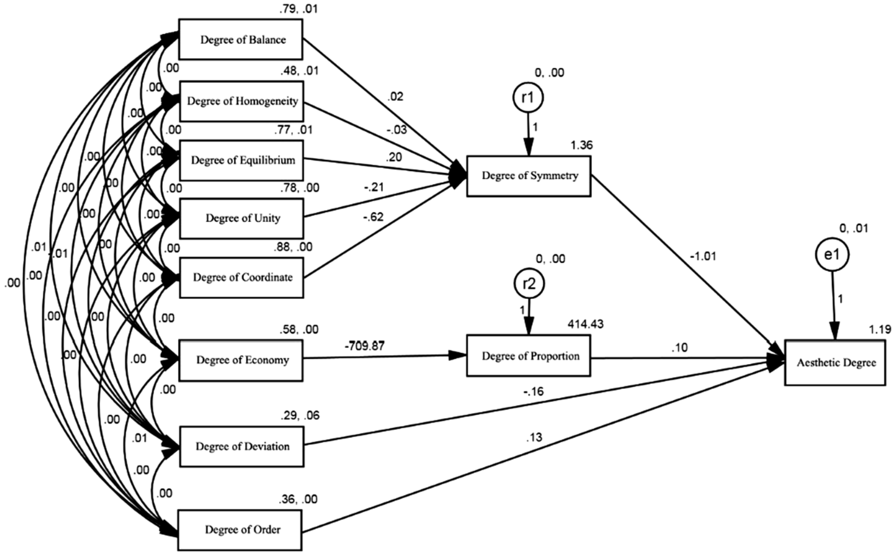


Fig. 3. The hypothesis model 2

4.3 Aesthetics Degree Evaluation System of Product Form

In the product form aesthetic evaluation system, the input parameter is the aesthetic degree indexes of the research sample, and the output parameter is the aesthetic image evaluation value after statistical analysis by the semantic difference method. After constructing the product form aesthetic evaluation system, it is necessary to carry out verification analysis. If the verification requirements are met, the system can effectively predict the product aesthetics. The BP neural network is essentially a nonlinear model structure, which is generally used to combine the complex relationship between input variables and output variables, and the BP network algorithm is the structure that fits the nonlinear function relationship. Compared with other networks, the advantages of establishing a model are relatively simple, do not require prior knowledge and rules for solving problems, have very good adaptability, and are suitable for predictive solutions of nonlinear models.

This study used BP neural network toolbox in MATLAB. The input layer parameters are 4 representative degree of aesthetic value, and the output layer parameters are aesthetic feeling by semantic differential in trained process, this network is set up 10 nodes in hidden layer, logsig-tansig transfer function, 400 maximum learning and 10^{-4} convergence error target. What this model trained used trainlm function and parameters in Table 2.

Table 2. Aesthetic degree indexes of samples and aesthetic feeling

Sample	Degree of symmetry	Degree of proportion	Degree of order	Degree of deviation	Aesthetic
1	0.7752	0.5000	0.3939	0.1118	0.8391
2	0.8231	0.7072	0.3328	0.3284	0.7609
3	0.8524	0.5000	0.3164	0.4187	0.7141
4	0.8328	0.5000	0.3230	0.3865	0.6828
5	0.8467	0.5774	0.3394	0.2280	0.5578
6	0.7844	0.7072	0.3007	0.4756	0.8469
7	0.8666	0.5000	0.3565	0.0052	0.6750
8	0.7774	0.6180	0.3943	0.1796	0.7922
9	0.7963	0.7072	0.2919	0.6449	0.7141
10	0.8102	0.7072	0.3493	0.4720	0.7297
11	0.8157	0.5000	0.3628	0.4834	0.6750
12	0.7531	0.5000	0.3959	0.1058	0.8938
13	0.7577	0.5000	0.3857	0.2080	0.7375
14	0.7780	0.7072	0.3564	0.1491	0.6359
15	0.7723	0.5000	0.3453	0.1619	0.6359
16	0.7914	0.5774	0.3792	0.0774	0.9094
17	0.8272	0.5000	0.3677	0.0588	0.8078
18	0.8017	0.5000	0.3428	0.9837	0.5813
19	0.7486	0.7072	0.3957	0.2611	0.7688
20	0.7832	0.5774	0.3840	0.0415	0.8625

After the BP neural network training of the teapot form aesthetic evaluation system is completed, the network should be verified and analyzed. In this study, five test samples were used to verify and compare the aesthetic evaluation system, and the representative aesthetic quantitative indicators of each verification sample were input into the system to predict the aesthetic evaluation values of the test samples, as shown in Table 3. Validation analysis for sample 1, sample 5, sample 9, sample 14, and sample 19, respectively. Through comparative study, only one of the five groups of data has a deviation greater than 0.125, and the network correct rate is 85%. Therefore, the verification of the purple sand pot shape aesthetic evaluation system using BP neural network is reasonable and acceptable.

Table 3. Aesthetic degree indexes of samples and aesthetic feeling

Sample	Aesthetic prediction	Aesthetic survey	Error
1	0.8353	0.8391	0.038
5	0.4918	0.5578	0.066
9	0.6821	0.7141	0.032
14	0.4689	0.6359	0.167
19	0.7398	0.7688	0.029

5 Construction Aesthetics Degree Optimization Prototype System

The genetic algorithm will represent a potential solution set of problems as a population consisting of a number of genetically encoded individuals, each of which is a possible solution to the problem, called a chromosome. Subsequent evolutionary calculations on these chromosomes are called genetic operations, and genetic operations are mainly achieved through three operations: selection, crossover, and mutation. The next generation of chromosomes obtained by crossing or mutating the selected chromosomes is called a progeny. Genetic algorithm application fitness measures the degree to which each individual in a group achieves an optimal solution in the operation. The fitness is used to measure the quality of the chromosome, and according to the degree of fitness, a certain number of individuals are selected from the previous generation and the offspring population as the father to continue to evolve. After several genetic operations, the algorithm converges to the chromosome with the highest fitness. The corresponding decoding is probably the optimal solution or the suboptimal solution to the problem. The fitness function is a function of measuring individual fitness, and its definition is generally different depending on the specific problem.

The operation content and basic steps that the genetic algorithm needs to complete are as follows:

First step, select the applicable encoding method to convert the problem parameter space into a string encoding space;

Second step, establish fitness function;

Third step, determining the genetic strategy mainly includes: population size, method of selecting operations, method of cross-operation, and method of mutation operation;

Fourth step, determine parameters such as crossover probability and mutation probability;

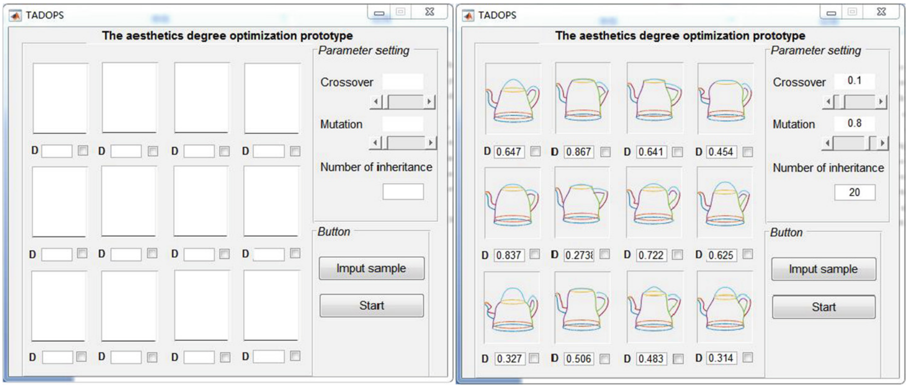
Fifth step, initialize the initial population;

Sixth step, calculate the fitness of each individual in the population;

Seventh step, according to the genetic strategy, three genetic operations of selection, crossover and mutation are carried out to obtain the progeny species;

Eighth step, Determine whether the population characteristics meet the expected requirements or have completed the predetermined number of iterations. If the above conditions are met, the operation ends and the result is output. Otherwise, return to step 7 or re-adjust the genetic strategy and return to step 7.

In this study, the spline curve was drawn by extracting the key points of product form in MATLAB, so the shape of the product was expressed. And the key points are binary coded, and the corresponding mutation and crossover probability are set. The fitness function is the above-mentioned product form aesthetic evaluation system, and finally the new purple clay teapot with higher aesthetic was gain. So back and forth, the innovative form will be found and the design method was developed. The aesthetics degree optimization prototype system (TADOPS) can provide a large number of product forms (Fig. 4). The TADOPS system was used to optimize the sample 8 and verified by 23 students with design background. The error rate was 26.6%, indicating that the optimization method based on product form aesthetics is practical and feasible.



Crossover 0.1

Mutation 0.8

Number of inheritance 20

D 0.647 □ D 0.867 □ D 0.641 □ D 0.454 □

D 0.837 □ D 0.273 □ D 0.722 □ D 0.625 □

D 0.327 □ D 0.506 □ D 0.483 □ D 0.314 □

Fig. 4. The aesthetics degree optimization prototype system

6 Discussion

In summary, we have presented a prototype system of teapot aesthetics optimization and introduced 10 aesthetic measures. The results shown the optimized products have obtained higher score which proved by 73.4% interviewees. This is probably a consequence of the aesthetic degree optimization model of product form simply works for morphology aesthetic, there are still some issues worthy of further analysis and discussion.

(1) The form of product is only considered in calculation of aesthetic degree, There are still research spaces for the three elements of morphology (form, color, texture), and their color and texture will affect human cognition of beauty.
(2) For the understanding of aesthetic, human have a very broad understanding. Morphology aesthetic, technical aesthetic, function aesthetic, artistic aesthetic and ecological aesthetic are judged form different angles. A lot of work is needed to enrich the aesthetics of the knowledge base.
(3) The connotation of beauty is changing, and the emergence of new and more complete things will surely lead to the re-evaluation of aesthetic standards by society. So corresponding aesthetic evaluation criteria will also change. In the future, the artificial intelligence and big data can be applied to establish a dynamic aesthetic evaluation system.

References

1. Nagamachi, M.: Kansei engineering as a powerful consumer-oriented technology for product development. Appl. Ergon. **33**(3), 289–294 (2002)
2. Birkhoff, G.D.: Aesthetic Measure. Cambridge University Press, Cambrige (1933)
3. Talia, L., Noam, T.: Assessing dimensions of perceived visual aesthetics of web sites. Int. J. Hum.-Comput. Stud. **60**(3), 269–298 (2004)
4. Schenkman, B.N., Jonsson, F.U.: Aesthetics and preferences of web pages. Behav. Inf. Technol. **19**(5), 367–377 (2000)
5. Ngo, D.C.L., Teo, L.S., Byrne, J.G.: Modelling interface aesthetics. Inf. Sci. **152**, 25–46 (2003)
6. Hsiao, S.W., Chou, J.R.: A Gestalt-like perceptual measure for home page design using a fuzzy entropy approach. Int. J. Hum.-Comput. Stud. **64**(2), 137–156 (2006)
7. Li, M., Zhang, J.: Research on aesthetics degree evaluation method of product form. In: Ahram, T., Karwowski, W., Taiar, R. (eds.) IHSED 2018. AISC, vol. 876, pp. 68–75. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-02053-8_11

Research on Evaluation of Product Image Design Elements Based on Eye Movement Signal

Wenjin Yang[1], Jianning Su[1,2(✉)], Kai Qiu[1], Xinxin Zhang[3], and Shutao Zhang[2]

[1] School of Mechanical and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, China
854415628@qq.com, sujn@lut.cn, qric@foxmail.com
[2] School of Design Art, Lanzhou University of Technology, Lanzhou 730050, China
[3] School of Art, Design and Media, East China University of Science and Technology, Shanghai 200237, China

Abstract. A product design elements evaluation model was proposed, which was constructed by eye tracking experiments, using eye movement indicators such as first gaze time, gaze order, and number of times of return, to accurately and effectively analyze product model and quantify users' emotion. The purpose of sorting the product model contribution was achieved through the weight calculation of the design elements. The soymilk machine shape was used as a case to verify the rationality of the evaluation model. The results showed that the "cup" part of the soymilk machine was the most concerned by the subjects, while the "handle" part was the least concerned. At the same time, the further research direction of design elements evaluation in eye movement experiment and the coupling between design elements, and product modeling image were discussed at the end of the article. The further research directions and experimental design methods were clarified.

Keywords: Product model design · Eye movement data · Design elements · Contribution ranking

1 Introduction

With the upgrading of the consumer market, product design gradually takes the user as the center instead of the product as the center, and develops in the direction of reflecting the user's emotional needs [1]. Product model is an important carrier to express product design ideas and disseminate spiritual and cultural significance [2]. The model of the product consists of design elements, including feature elements and platform elements. The feature elements are the main emotional demand representation of the consumer when selecting and using the product. Therefore, designers can quickly and accurately unearth the needs of consumers by sorting the contribution of product design elements to complete product design. At present, in the evaluation stage

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 214–226, 2019.
https://doi.org/10.1007/978-3-030-22507-0_17

of product design elements, subjective questionnaires are combined with data calculation [3], the little research was directed at systematic objective evaluation.

The temporal and spatial characteristics of eye movements are the physiological and behavioral manifestations in the process of visual information extraction. As a feedback of physiological awakening, it has been widely used in the fields of psychology and medicine. Judging the psychological state of consumers based on visual trajectory has become a hot topic in the field of product design, Li et al. [4] established a mathematical model between subjective scoring and multiple eye movement data, which solved the problem of poor reliability of subjective selection results in product modeling schemes and lack of comprehensive quantitative research on objective selection results. Tang et al. [5] conducted a correlation analysis between EEG, eye movement data and subjective evaluation values, and established a program of car user experience selection model combining physiological and psychological evaluation indicators; At the same time, physiological data is playing an increasingly important role in the identification and evaluation of elements, Kristian et al. [6] search for the best combination of packaging elements through orthogonal design and physiological measurement; Cheng et al. [7] unearthed user preferences in interface design by using eye tracking experiments. The above eye movement physiology data is mainly used to evaluate the product plan, and the method used in the extraction evaluation study of the design element is subjective and lacks of quantitative method verification [8].

An evaluation model of product design elements contribution driven by eye tracking data is proposed. The physiological awakening amount of the user is used as the evaluation basis, and the eye movement index calculation data is used as the evaluation result of the design elements weight, and the product design elements contribution ranking is obtained to guide product design scientifically and quantitatively.

2 Unearthing the Target Image of Product Model

Target image often represents the user's main evaluation of product model. At present, the interview method, the oral analysis method, the image scale method, etc. are used by the main tester to explore the image of the product model target. Entropy is used as an important indicator to measure system stability, information entropy [9] is a negative entropy used to indicate the degree of order of the system, which can be used to theoretically construct a product modeling image evaluation model based on the composite cognitive space of users, designers and engineers to guide the selection of typical case libraries [10]. The principle of information entropy is used as the method to unearth the most representative product model image, as the theoretical basis of eye tracking experiment.

Product image evaluation value characterized by entropy:

$$I_j = -k \sum_{i=1}^{m} F_{ij} \ln F_{ij} \tag{1}$$

In the formula: I_j represents the image entropy value; F_{ij} represents the jth target image probability of the i-th sample; i represents the research sample, $i = 1, 2, \ldots, m$; j represents the target image, $j = 1, 2, \ldots, q$; k represents a constant, $k = 1/\ln m$.

Through the semantic differential method, the target image probability is obtained after normalization. Applying the formula (1) to obtain the entropy value I_j of the target image of the jth item, the weight of the target image in the evaluation process is w_j:

$$w_j = \frac{1 - I_j}{\sum_{j=1}^{n} (1 - I_j)} \tag{2}$$

According to the weight value of each image, the largest as the target image will be chosen.

3 Visual Cognition of Product Imagery

(1) Visual cognition

Visual cognition refers to the physiological response of the eye after receiving external stimuli. Through the stimulation of product styling, the image information is transformed into a visual signal and transmitted for the brain to form a psychological stimulus. The whole process is the mutual confirmation of physiological response and psychological stimulation, which is shown in Fig. 1. It can be summarized as overall organization, constant memory, simple adjustment and selection resolution [11]. The indicators mainly include physiological parameters such as gaze, follow, and saccade. The visual tracking experiment focuses on the average pupil diameter, gaze time, gaze order, and number of times of return.

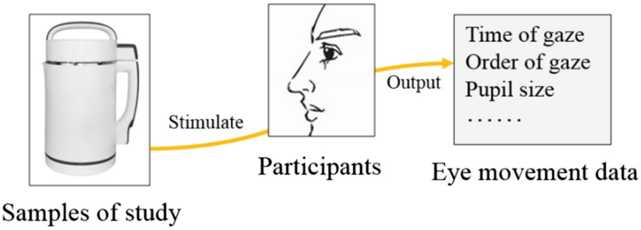


Fig. 1. Schematic diagram of visual cognition process

Assume that the subjects are gathered during the experiment as $P = \{p_1, p_2, \ldots, p_n\}$, the sample set is $S = \{s_1, s_2, \ldots, s_m\}$, various eye movement indicator data sets $E = \{e_1, e_2, \ldots, e_k\}$. Then, the product image design elements evaluation experiment data set is:

$$Y = \begin{bmatrix} y_{(p_1,s_1,E)} & y_{(p_1,s_2,E)} & \cdots & y_{(p_1,s_m,E)} \\ y_{(p_2,s_2,E)} & y_{(p_2,s_2,E)} & \cdots & y_{(p_2,s_m,E)} \\ \cdots & \cdots & \cdots & \cdots \\ y_{(p_n,s_1,E)} & y_{(p_n,s_2,E)} & \cdots & y_{(p_n,s_m,E)} \end{bmatrix} \quad (3)$$

Where $y(p_i, s_h, E)$ represents the eye movement index data of the subject s_h to the subject $p_{i.}$

(2) Eye movement index

① Average diameter of the pupil
It refers to the average diameter of the pupil, which is mainly used to detect the psychological activity of the sample to be observed. The size of the pupil diameter is related to the mood of the subject. Under normal circumstances, the pupil is enlarged when the mood is high, and the pupil is reduced when the mood is low [12].

② Time of gaze
It refers to the time difference between the first time the subject is gazing at the departure of a certain area of the sample. Combined with the eye movement data analysis software, the visual center distribution of the subject during the experiment can be understood.

③ Order of gaze
It refers to the order in which the subjects look at all areas of the sample during the experiment. The analysis of the data can indirectly obtain the subject's interest in all areas of the sample.

④ Number of times of review
It refers to the number of times the subject repeatedly looks at the local area after the sample is observed and re-watches the area that has already been watched. It represents the degree of which the subject pays attention to the area being looked back.
In the course of this experiment, because the test time is short, it can't cause the subject's emotional fluctuations. Therefore, the gaze time, the gaze order, and the number of times of return are used as the basis for evaluation and analysis.

(3) Eye movement experiment

The general flow of eye movement experiment is shown in Fig. 2.

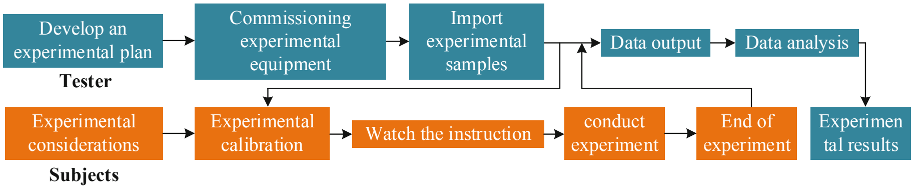


Fig. 2. Eye movement experiment flow chart

① The main tester formulates the eye movement experiment plan and tasks, pre-commissions the eye movement experiment equipment, etc., and introduces the designed experimental materials into the eye movement experiment software system.
② The main tester explained the purpose of the experiment and the precautions to the subject, and obtained the consent of the subject.
③ Perform eye movement test calibration on the subject under the guidance of the main tester.
④ After confirming the calibration, perform the experimental phase. First, the subject should watch the instruction, then press "Space" or "→" to enter the next page until the experiment is completed. During the experiment, the subject can make appropriate visual adjustments or rest according to the requirements, this process is not included in the statistics of the research results.

4 Product Design Elements Evaluation Model Under Target Image

According to the literature [13], the subject's gaze time for the design element g is defined as t, Where x is the order in which the subject is gazing; v_1 is a sequence factor that distinguishes the importance of the order, and its value ranges from [1, 2]. When the first gaze, the gaze factor takes a value of 2. The elements are sequentially reduced by x/u after each gaze; u is the number of all effective fixation points for each test sample experiment, and the gaze time $t(2 - x/u)$ of each design element g is obtained according to the order of gaze.

When a design element has a time of return t' in the test, the number of times the subject has returned to the design element g is defined as α. v_2 is the review factor in the experimental process, and the value range is [1, 2], wherein the value of the first back view is 1, and as the number of review increases, the review factor is gradually increased by 1/20. Each design element g may have multiple views, so the return time of each design element g should include the sum of multiple back time, and each time $t'(1 + (\alpha - 1)/20)$ of each design element is obtained.

Finally, the weight of each design element g is

$$weight = \frac{1}{mn}\sum_{h=1}^{n}\sum_{i=1}^{m}\frac{t_{(i,h)}(2 - \frac{x}{u}) + \sum_{\alpha=1}^{\beta} t'_{(i,h)}(1 + \frac{\alpha-1}{20})}{total} \quad (4)$$

In the formula: *Weight* represents the weight value of a design element, which is ranged from [0, 1]; i represents the research sample, i = 1, 2, …, m; h represents the subject, h = 1, 2, …, n; β represents the total number of times the design element g is viewed back; Total represents the sum of all valid gaze times for completing a sample experiment.

In the ranking contribution of the whole design elements, the higher the front, the greater the weight value, which indicated that needs to be valued in the product image design.

5 Case Study

5.1 Identify Research Samples and Initial Emotional Images

As an experimental research sample, 85 initial samples of soymilk machine were collected from websites, periodicals, newspapers and other fields totally. Through expert discussion, similar samples were removed, and 15 final research samples were obtained. In order to eliminate interference from other factors, remove the elements such as sample color and logo, then obtain the research sample set S = {s_1, s_2, …, s_{15}}, as shown in Fig. 3. Using network data survey and cluster analysis, the experimental sample modeling image set B = {Atmospheric, Succinct, Exquisite, Stylish} is obtained.

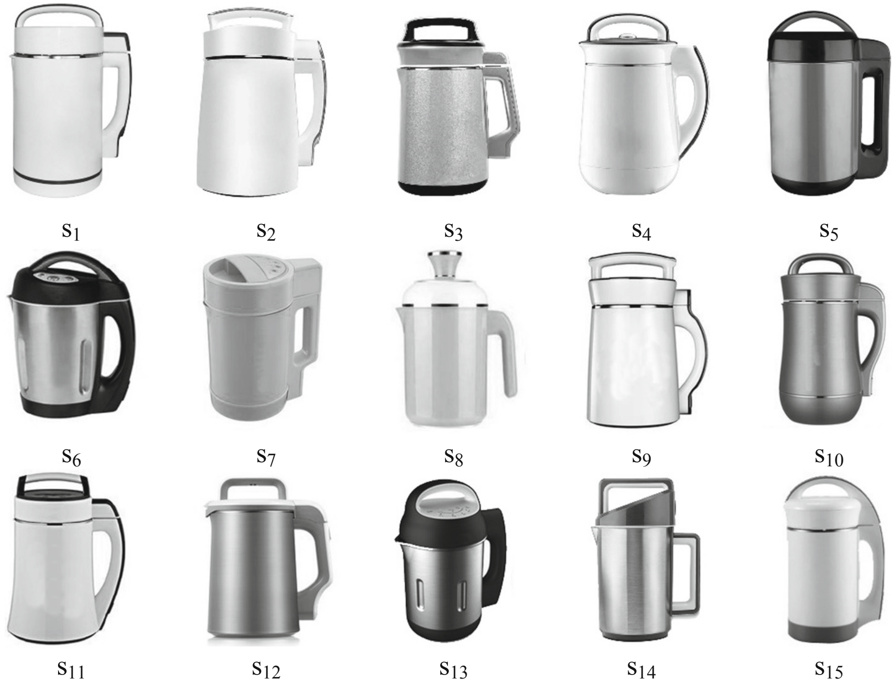


Fig. 3. Research sample set

5.2 Identify the Target Image

Combine the sample dataset with the modeling image set to create a SD 5 rating questionnaire, in the form of semantic difference method, 14 postgraduates and 2 undergraduates in product design are surveyed, take the "atmosphere" as an example. The most atmospheric feeling is 5 points, and the least atmospheric feeling is 1 point. After the questionnaire was completed, the interviewees were interviewed by oral interview. The invalid questionnaire was removed, and 15 results were obtained. After the average value was processed, the formula (2) was used to calculate the weight of the modeling image as shown in Table 1.

Table 1. Weighted evaluation form of Sensual image

Sensual image	Image weight
Atmospheric	0.1976
Succinct	0.1978
Exquisite	0.2609
Fashionable	0.3437

According to the data in the table, the "Fashionable" has the largest weight, which is 0.3437. Therefore, the Stylish will be taken as the target image of the research sample in the subsequent research.

5.3 Specific Experimental Process

According to the sample picture, it is made into eye tracking research material, the purpose of evaluating the sample design elements is achieved through the eye tracking experiment. Taking sample 1 as an example, the experimental material 1 is obtained as shown in Fig. 4.

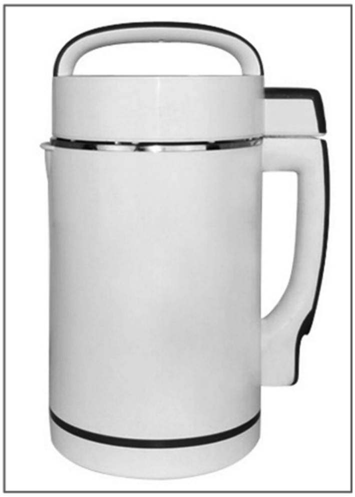


Fig. 4. Experimental material 1

(1) Set the experimental instruction: press "→" on the keyboard to enter the next page, use the "fashionable" as the evaluation basis, and score 5 points on the sample picture. The most fashionable is 5 points, the least fashion is 1 point, press "→" Go to the next page.

(2) Commissioning of experimental equipment by the main tester.

(3) Adjust the image material resolution to 1366 * 768 pixels and import it into the ErgoLAB software device along with the instructions. In order to reduce the influence of the previous sample evaluation, the test results were credible, and the set samples appeared randomly, but all of them appeared only one time, as shown in Fig. 5. Test the entire experimental process, observe the experimental environment, and enter the experimental stage after confirmation.

(4) Subjects

The purpose of this experiment is to evaluate the sample of the study based on the image, and to achieve the purpose of evaluating the design elements through the calculation of eye movement data. In this experiment, there were 13 students in the school, including 2 undergraduates, 10 postgraduates, and 1 doctoral student. All the subjects were design-related majors, and they had a certain understanding of the Kansei Engineering. They were familiar with the research and development of the soymilk machine in terms of function and model. Through the explanation of the main tester, the subjects indicated that the experiment could be successfully completed, and all subjects had corrected visual acuity of 1.0 or above, right hand.

(5) Eye movement calibration

The main tester informed the subject to start the calibration test before the experiment, the first is sitting posture and distance adjustment. The two spots are presented in the software interface after the eyes of the subject were captured by the eye tracker as shown in Fig. 6. When the two strips on the right and the bottom are stably green, the eye tracker can detect the visual information to the subject and the signal is stable.

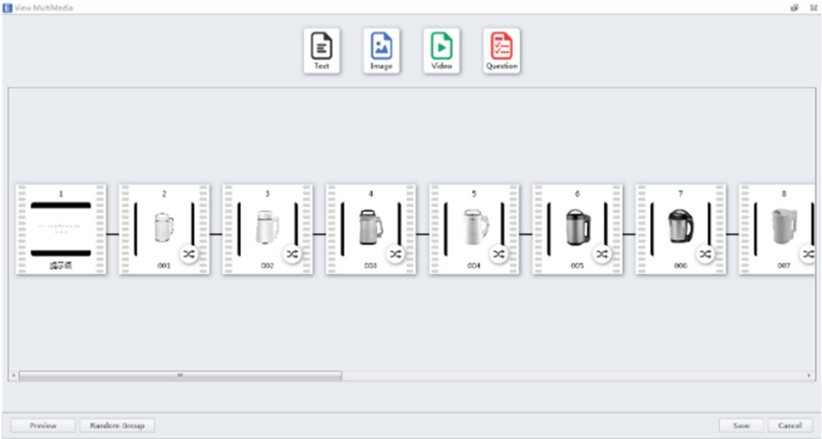
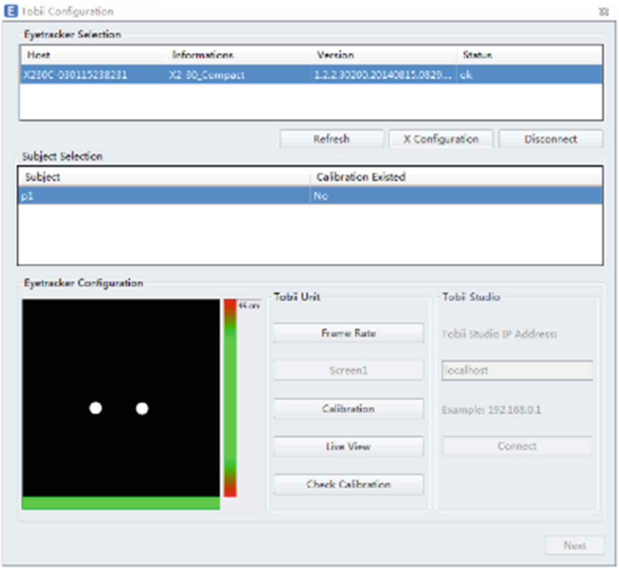


Fig. 5. Map of experimental material import



Host	Informations	Version	Status
X230C-030115238231	X2-30_Compact	1.2.2.30200.20140815.0829...	ok

Subject	Calibration Existed
p1	No

Tobii Studio IP Address: localhost

Example: 192.168.0.1

Fig. 6. Calibration interface for eye movement experiments

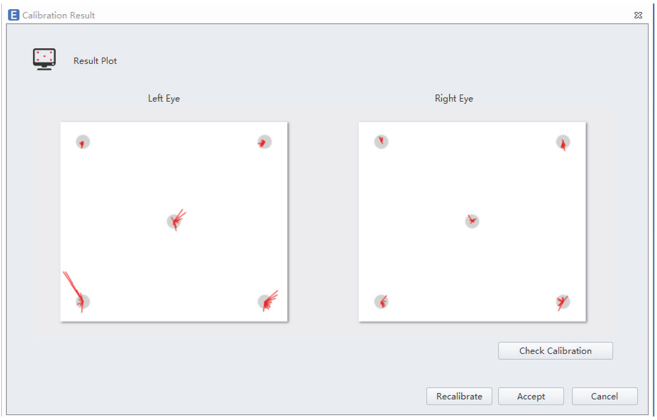


Fig. 7. Eye movement calibration

Next process is the calibration, the subject will see a moving red dot, the line of sight follows the point of movement, the result after the end of the calibration as shown in Fig. 7, showing the error vector of the calibration, each the size of the red line segment represents the difference between the calculated gaze point and the actual rendered calibration point. If the red line segment of a point is long or there is no line segment, it needs to be rescaled. After the calibration is completed, start the experiment.

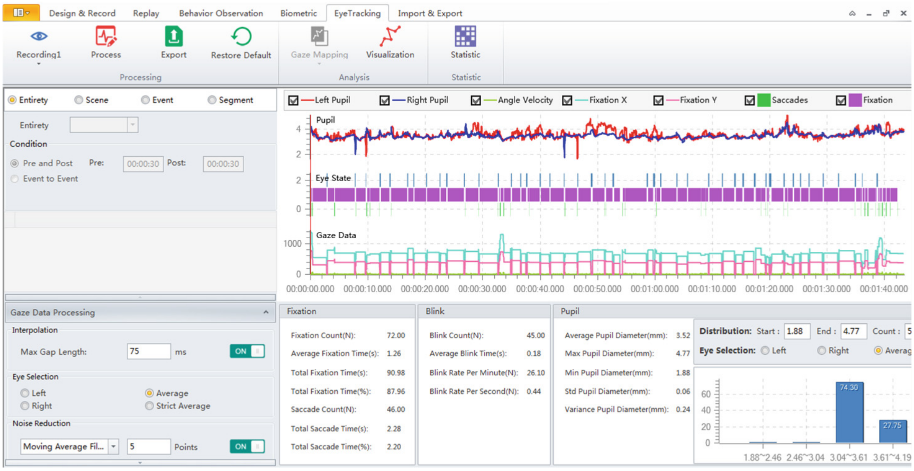
(6) After the preparation of the pre-experiment, the main tester guides each subject to sit in the fixed seat and informs the purpose of the experiment, which is to

calculate the eye tracking data through analysis and to achieve the purpose of the product elements evaluation.

(7) 13 subjects underwent eye tracking experiments in sequence until the end.

(8) Statistical analysis of experimental results

The experimental data output interface as shown in Fig. 8. The data was statistically analyzed by ErgoLAB software. According to the purpose and requirements of the experiment, unreasonable experimental data is removed (eye movement information is not on the sample or eye tracker does not collect eye movement information, etc.), the statistical data includes the subject's first gaze time, number of gaze, and time of return, and the order of viewing, etc.



● Entirety ○ Scene ○ Event ○ Segment

Entirety

Condition

● Pre and Post Pre: 00:00:30 Post: 00:00:30

○ Event to Event

Interpolation

Max Gap Length: 75 ms ON

Eye Selection

○ Left ● Average

○ Right ○ Strict Average

Noise Reduction

Moving Average Fil... 5 Points ON

Fixation

Fixation Count(N):	72.00
Average Fixation Time(s):	1.26
Total Fixation Time(s):	90.98
Total Fixation Time(%):	87.96
Saccade Count(N):	46.00
Total Saccade Time(s):	2.28
Total Saccade Time(%):	2.20

Blink

Blink Count(N):	45.00
Average Blink Time(s):	0.18
Blink Rate Per Minute(N):	26.10
Blink Rate Per Second(N):	0.44

Pupil

Average Pupil Diameter(mm):	3.52
Max Pupil Diameter(mm):	4.77
Min Pupil Diameter(mm):	1.88
Std Pupil Diameter(mm):	0.06
Variance Pupil Diameter(mm):	0.24

Distribution: Start : 1.88 End : 4.77 Count : 5

Eye Selection: ○ Left ○ Right ● Averag

Fig. 8. Output interface of eye movement data

5.4 Output Experiment Results

After all the subjects completed the experiment in turn, the main tester used the ErgoLAB eye movement data processing software to output the eye movement data, and the output content included the subject's gaze time, blink frequency, pupil change, eye movement heat map and the like. The main tester conducted statistical analysis on the data and obtained experimental conclusions.

The soymilk machine model is divided into: handle (g_1), nose (g_2), cup (g_3) and grip (g_4) through the morphological analysis method and expert discussion method, taking sample 1 as an example, the distribution of design elements as shown in Fig. 9. According to the calculation of formula (4), the proportion of each design element of the sample is obtained, as shown in Table 2.

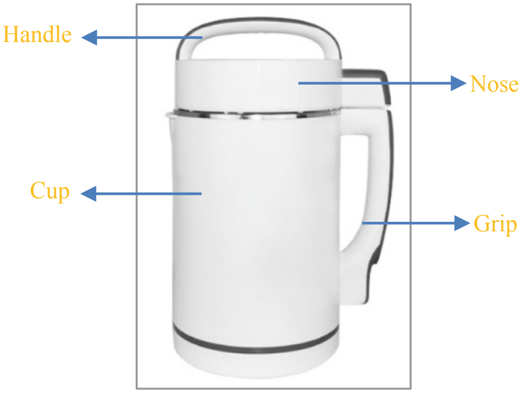


Fig. 9. Distribution map of design elements

Table 2. Weights of each design element

Design elements	handle (g_1)	nose (g_2)	cup (g_3)	grip (g_4)
Weight data	0.2092	0.9797	0.9881	0.4166

6 Discussion

Eye movement information is a representation of physiological signals, as well as a reflection of psychological cognition. It is a driving force for the evolution of product model. In the course of our interview, all the subjects said that when evaluating the product model image, the first concern is the overall shape of the sample. After having a preliminary grasp of the whole, we will consider the details of the product, it is the evaluation of product design elements. Among them, 84.6% of the subjects indicated that the focus was the cup, and 69.2% of them indicated that the second was the nose, which is consistent with the calculated results. At the same time, relevant research shows that the consumer's psychological cognition and physiological response are consistent [14], so the eye movement tracking product design elements evaluation model is reasonable.

Of course, there may be other reasons for this. When we interviewed other designers, including the subjects, part of them thought that in almost every sample, the area of the "cup" would be larger than other design elements, which would make the eye movement information fall in the area. The chance of being physically will be significantly higher than other design elements, resulting in the presentation of results. Similarly, part of them said that during the experiment, the cup is often in the middle of the screen, which will first attract visual attention. It will mislead to such experimental results.

In the follow-up study, we will divide into two experiments. The first is to explore the correlation between area and eye movement information. The area of the design element is used as an independent variable, and the eye movement information is used as a dependent variable. The second part is to randomly change the position of the sample on the screen to explore the relationship between the position of the design elements and the eye movement information. Combining the above two experimental results, the rationality of eye movement information and design factor evaluation model is discussed again.

Product design element is an important part of product model. In terms of function, any design element is indispensable. In terms of form, the design elements and the correlation between them constitute the overall model of the product, and there is a coupling relationship between them. Generally, we think that the more coupling between the elements, the better, because it represents the integrity of the product model, but the degree of coupling is too high, which often makes the product lack of uniqueness, so that it loses the beauty of model design. What is the optimal range of coupling between design elements to ensure the integrity and aesthetics of the product model? Therefore, the evaluation of the coupling degree between product design elements and the correlation include the coupling between design elements and product model images will be the hot issues of subsequent research.

Acknowledgments. The authors would like to extend our sincere thanks to all those who contributed to this study, especially the subjects who took the trouble to take part in experiments. The project is sponsored by National Natural Science Foundation of China (51465037), Hongliu Outstanding Talent Development Program of Lanzhou University of Technology (J201406).

References

1. Chen, C.H., Sato, K., Lee, K.P.: Editorial: human-centered product design and development. Adv. Eng. Inform. **23**(2), 140–141 (2009)
2. Su, J.N., Wang, P., Zhang, S.T., et al.: Research progress on key technologies of product image model design. J. Mech. Des. **30**(1), 97–100 (2013)
3. Niu, W.P., Yao, J., Li, X.Z., et al.: Interior design of medium-sized excavator facing to driver security needs. J. Mech. Des. **33**(7), 125–128 (2016)
4. Li, Y., Guo, G.: Selection model of product shape schemes based on multiple eye movement data. J. Comput. Integr. Manuf. Syst. **22**(3), 658–665 (2016)
5. Tang, B.B., Guo, G., Wang, K., et al.: User experience evaluation and selection of automobile industry design with eye movement and electroencephalogram. J. Comput. Integr. Manuf. Syst. **21**(6), 1449–1459 (2015)
6. Kristian, P., Tanel, M., Andres, K.: Considering emotions in product package design through combining conjoint analysis with psycho physiological measurements. Procedia Soc. Behav. Sci. **148**(25), 280–290 (2014)
7. Cheng, S.W., Liu, Y.: Eye-tracking based adaptive user interface: implicit human-computer interaction for preference indication. J. Multimodal User Interfaces **5**(2), 77–84 (2012)
8. Wang, Z.Y., Li, H.W.: Automotive styling feature extraction and cognition based on the eye tracking technology. Packag. Eng. (20), 54–58 (2016)
9. Gray, R.M.: Entropy and Information Theory, 2nd edn. Springer, NewYork (2011). https://doi.org/10.1007/978-1-4419-7970-4
10. Su, J.N., Zhang, X.X., Jing, N., et al.: Research on the entropy evaluation of product styling image under the cognitive difference. J. Mech. Des. **33**(3), 105–108 (2016)

11. Koffka, K.: The Principle of Gestalt Psychology. Zhejiang Education Publishing House, Hangzhou (1997)
12. Antonio, L., Gaetano, V., Enzo, P.S.: Eye gaze patterns in emotional pictures. J. Ambient Intell. Humaniz. Comput. **4**(6), 705–715 (2013)
13. Chen, X.J., Yan, H.X., Xiang, J.: Study of decoding mental state based on eye tracks using SVM. Comput. Eng. Appl. **47**(11), 39–42 (2011)
14. Zhang, X.W.: Research of modeling theory and method in computational psychophysiology. Lanzhou University (2016)

Measurement of Human Sitting Posture Dimensions Using Human Pressure Distribution

Chen Yue[1], Linghua Ran[2](✉), and Hua Qin[1]

[1] Department of Industrial Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

[2] China National Institute of Standardization, No. 4 Zhi Chun Road, Haidian District, Beijing 100191, China

ranlh@cnis.gov.cn

Abstract. The BIOFORCEN pressure distribution test system was used to measure the body's sitting position dimensions which were important for the seating products design, including the sciatic distance, the distance from the edge of the ischial bone to the lateral side of the hip, and the distance from the posterior to the hip to the back of the hip. And at the same time the Martin instrument was used to measure body's basic items which included the body height, weight, sitting hip width, sitting height, hip and hip distance of the subject. The sitting position dimensions were compared with the data of the United Kingdom and the United States. And the correlation of the sitting position dimensions and the basic body items were analyzed and the regression equations were established.

Keywords: Sitting posture dimensions · Pressure distribution · Correlation

1 Introduction

In daily life, from clothing, seats to mattresses, the contact and friction between human body and various objects will exert certain pressure on the body [1]. Sitting for a long time has increasingly become the prime culprit of many office workers' chronic diseases [2], and excessive local pressure or improper overall pressure distribution have a great negative impact on the human body. If the local pressure on the hip is too heavy or too long, it will block the microvascular blood circulation of the pressed part, affect nerve endings and cause numbness or pain [3, 4]. Good seats don't cause fatigue easily [5–7].

For seats comfort, subjective and objective evaluation methods are used by researchers, and the research direction is mainly focused on seat materials, seat design parameters and seat comfort evaluation [8–11]. Jiang-hongzhao, tang, etc. in 1991, has been using the method of combining subjective and objective evaluation to study the static comfort of passenger seat back curve. For the objective evaluation the research focused in three aspects: the curve of human body measurement, electrical measurement, posture and action observation research [12–16]. Qun-sheng and xia put forward

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 227–235, 2019.
https://doi.org/10.1007/978-3-030-22507-0_18

eight characteristics of seat comfort index of body pressure distribution [15–18]. Zhang Zuyin put forward the basic principles of seat design. By using the principles of human anatomy and the posture behavior analysis. Song haiyan et al. studied the effect of sitting height on the comfort of office seats by using subjective evaluation method, and the results showed that the seat was most comfortable when the knee height was higher 2.14 cm than the seat surface height [19–22].

The human body dimensions are important to the seats design. Some basic sitting position dimensions, such as sitting height, sitting breadth and so on can be measured by traditional methods. But some dimensions cannot be directly accessible due to technical constraints, such as the sciatic distance which will be hidden by chairs when people sit down [23]. When the person sits down, the pressure of the human sciatic tubercle in contact with the seat is the greatest [24]. At this point, the area with the greatest pressure on the pressure pad is the location of the ischial tubercle. This paper tried a new method to use pressure cushion to test these data which are not convenient to be measured by traditional methods [25]. These data include the sciatic distance, the distance from the edge of the ischial bone to the lateral side of the hip, and the distance from the posterior to the hip to the back of the hip.

2 Data Collection Experiment

The experiment was conducted on November 2, 2018 at the China National Institute of Standardization. The experiment tested a total of 18 people, 9 men and 9 women. The ratio of men and women was balanced. The subjects with different BMIs were evenly distributed. Normal, fat and thin each accounted for 6 people. The personal information is in the Table 1.

Table 1. The personal information

Number	Gender	Age	Height (cm)	Weight (k)	BMI	BMI sort
1	m	22	185	74.2	21.68006	Normal
2	m	24	170	59.8	20.69204	Normal
3	m	22	172	68	22.9854	Normal
4	m	23	171.3	87.7	29.88718	Obesity
5	f	23	165.4	57.5	21.01826	Normal
6	f	21	160	54	21.09375	Normal
7	f	32	161	54	20.83253	Normal
8	f	32	161.4	45.7	17.54321	Thin
9	f	46	153.6	62	26.27903	Obesity
10	f	25	154.2	63.2	26.57959	Obesity
11	f	32	155.3	70.8	29.35556	Obesity
12	m	42	172	87.4	29.543	Obesity
13	m	23	185	104	30.38714	Obesity
14	f	21	156	45	18.49112	Thin

(*continued*)

Table 1. (*continued*)

Number	Gender	Age	Height (cm)	Weight (k)	BMI	BMI sort
15	f	22	166.6	48.5	17.47398	Thin
16	m	20	176.5	54.7	17.55892	Thin
17	m	21	176	55	17.75568	Thin
18	m	21	170	53	18.3391	Thin

During the test, the user was asked to sit on a hard chair with a pressure cushion on it. The user's legs were adjusted to be vertical to the ground, and the thighs to be parallel to the ground. The subjects sat in the chair and remained motionless for three seconds. The two points with the greatest pressure on the buttock and seat contact surface were obtained as the data of the ischial tubercle, and three data related to the ischial tubercle were measured.

At the same time, the hip width, sitting height, hip circumference were also been measured by the Martin measurement tools.

3 Data Analysis

3.1 Statistical Data

The statistical data results of the sitting position dimensions are shown below (Tables 2 and 3).

Table 2. Basic statistical data for male

	Isotopic spacing (mm)	Sciatic edge to the lateral side of the hip (mm)	Sciatic edge to the back of the hip (mm)
Mean	111.6	104.8	83.0
Standard deviation	17.79	16.3	28.5
Percentiles 5%	98.0	85.7	36.7

Table 3. Basic statistical data for female

	Isotopic spacing (mm)	Sciatic edge to the lateral side of the hip (mm)	Sciatic edge to the back of the hip (mm)
Mean	127.9	88.4	80.3
Standard deviation	10.8	8.1	16.3
Percentiles 5%	122.	85.7	61.2

3.2 Data Comparison Analysis

Comparative Analysis of Male Data

The three data of United States and the United Kingdom were obtained from literatures. The data are showed in the following tables (Table 4).

Table 4. Sciatic spacing

Item	Country	Sex	Mean	sd	5th %ile	95th %ile
Isotopic spacing	UK	m	117.9	11.3	99.4	136.5
		f	129.4	10.1	112.7	146.1
	USA	m	118.3	11.4	99.5	137.1
		f	129.9	11.1	111.7	148.1
Sciatic edge to the lateral side of the hip	UK	m	128.8	13.6	106.4	151.2
		f	137.2	12.6	116.5	158
	USA	m	130.4	15.5	104.9	156
		f	140	17.6	111	169
Sciatic edge to the back of the hip	UK	m	71.3	7.9	58.3	84.2
		f	74.7	9.6	58.9	90.4
	USA	m	72.2	9	57.4	87
		f	76.4	14	53.2	99.4

Comparison of US, UK and China data male mean values is shown in Fig. 1.

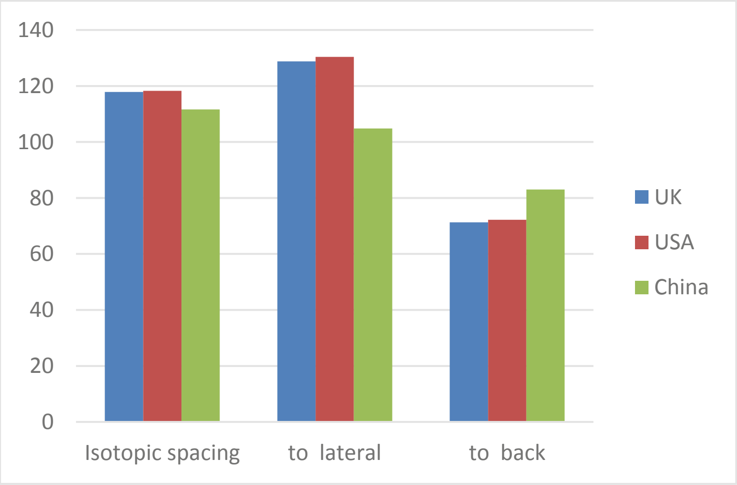


Fig. 1. Comparison of male means

From the above figure, the mean spacing of male ischial bones is 111.6 mm for Chinese male, which is smaller than the average of 117.9 mm for British men and 118.3 mm for American men. The average distance between the lateral edge of the male ischial and the lateral aspect of the buttock is 104.8 mm, which is smaller than the mean of the British male 128.8 and the average of the male of 130.4 mm. The average distance from the edge of the male ischial to the posterior hip is 83 mm. The mean value of the British male is 71.3 mm and the average of the American male is 72.2 mm.

Comparative Analysis of Female Data
Comparison of US, UK and China data female mean values is shown in Fig. 2.

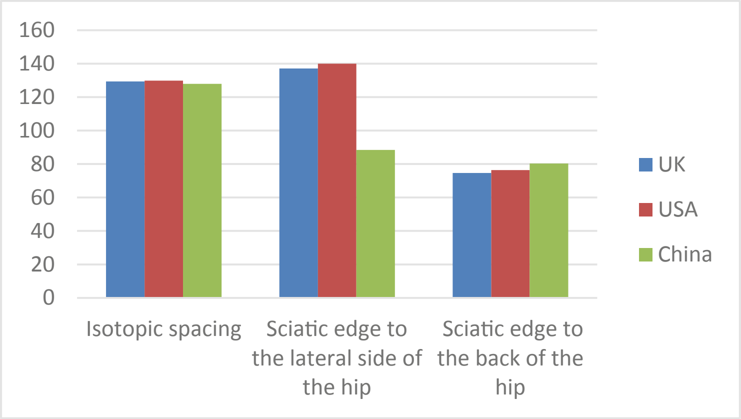


Fig. 2. Comparison of female means

From the above figure, the mean spacing of male ischial bones is 127.9 mm for Chinese female, which is smaller than the average of 129.4 mm for British men and 129.9 mm for American men. The average distance between the lateral edge of the male ischial and the lateral aspect of the buttock is 88.5 mm, which is smaller than the mean of the British female 137.2 mm and the average of the American female of 140 mm. The average distance from the edge of the male ischial to the posterior hip is 80.3 mm. The mean value of the British male is 74.7 mm and the average of the American male is 76.4 mm.

3.3 Correlation and Regression Analysis

Male Single Independent Variable Linear Correlation
Using the bivariate correlation analysis of SPSS software, the data in Table 5 can be obtained. The content shown in the table is the correlation coefficient and significance of the two variables with significant correlation.

Table 5. Male correlation analysis result

Correlation between values	Isotopic spacing (mm)	Sciatic edge to the lateral side of the hip (mm)	Sciatic edge to the back of the hip (mm)
Age	0.43	0.545	0.267
Height	0.548	0.739	0.392
Weight	0.011	0.046	0.013
Hip width (mm)	0.08	0.014	0.103
Sitting height (mm)	0.469	0.562	0.209
Hip circumference (mm)	0.825	0.354	0.199

Male Linear Regression Analysis

From the figure below, we can see data with significant correlation and formulas between them (Table 6).

Table 6. Correlation and formula for male

Number	Correlation	Formula
1	Separation distance, weight	Separation distance (mm) = 0.778 * weight (kg) + 55.967
2	Distance from the edge of the ischial bone to the outside of the buttocks, weight	Distance from the edge of the ischial bone to the outside of the buttocks (mm) = 0.608 * weight (kg) + 61.292
3	Distance from the edge of the ischial bone to the outside of the buttocks, sitting posture hip width	Distance from the edge of the ischial bone to the outside of the buttocks (mm) = 0.332 * sitting posture hip width (kg) − 16.515
4	Distance from the edge of the ischial bone to the back of the buttocks, weight	Distance from the edge of the ischial bone to the back of the buttocks (mm) = 1.235 * weight (kg) − 5.335

3.4 Analysis of Female Correlation and Regression Relationship

Female Single Independent Variable Linear Correlation

Using the bivariate correlation analysis of SPSS software, the data in Table 7 can be obtained. The content shown in the table is the correlation coefficient and significance of the two variables with significant correlation.

Table 7. Female correlation analysis result

Correlation between values	Isotopic spacing (mm)	Sciatic edge to the lateral side of the hip (mm)	Sciatic edge to the back of the hip (mm)
Age	0.829	0.657	0.118
Height	0.138	0.415	0.367
Weight	0.454	0.054	0.02
Hip width (mm)	0.593	0.082	0.165
Sitting height (mm)	0.014	0.701	0.661
Hip circumference (mm)	0.593	0.082	0.011

Female Linear Regression Analysis. From the figure below, we can see Data with significant correlation and formulas between them (Table 8).

Table 8. Correlation and formula for female

Number	Correlation	Formula
1	Distance from the edge of the ischial bone to the back of the buttocks, weight	Distance from the edge of the ischial bone to the back of the buttocks (mm) = 1.413 * weight (kg) + 1.684
2	Separation distance, sitting height	Separation distance (mm) = 0.475 * sitting height (mm) − 278.95
3	Distance from the edge of the ischial bone to the back of the buttocks, hip circumference	Distance from the edge of the ischial bone to the back of the buttocks (mm) = 2.701 * hip circumference + 20.675

4 Conclusion

The paper uses the BIOFORCEN pressure distribution test system for objective stress test. The system measures the sciatic distance of the participants, the distance from the edge of the ischial bone to the lateral side of the hip, and the distance from the posterior to the hip to the back of the hip. Data were analyzed and analyzed by spss software. The mean distance, standard deviation, and percentile of the distance between the ischial bone, the distance from the edge of the ischial bone to the lateral side of the scia, and the posterior to the sac of the scia were obtained. Comparing with the US and UK data, it is found that the variables with significant correlation, and the correlation formula were established.

The experiment draws the following conclusions:

(1) For men, the distance between the sciatic bone and the body weight, the distance from the edge of the ischial bone to the outside of the hip and the body weight, the distance from the edge of the ischial bone to the outside of the hip and the width of the sitting hip, the distance from the edge of the ischial bone to the back of the hip are strongly correlated with the body weight. Get their correlations as follows:
Separation distance (mm) = 0.778 * weight (kg) + 55.967
Distance from the edge of the ischial bone to the outside of the buttocks (mm) = 0.608 * weight (kg) + 61.292
Distance from the edge of the ischial bone to the outside of the buttocks (mm) = 0.332 * sitting posture hip width (kg) − 16.515
Distance from the edge of the ischial bone to the back of the buttocks (mm) = 1.235 * weight (kg) − 5.335
(2) For women, the distance from the edge of the ischial bone to the back of the buttock and the weight, the distance between the sciatic and the sitting height, the distance from the edge of the ischial bone to the back of the hip are strongly correlated with BMI. Get their correlations as follows:
Distance from the edge of the ischial bone to the back of the buttocks (mm) = 1.413 * weight (kg) + 1.684
Separation distance (mm) = 0.475 * sitting height (mm) − 278.95
Distance from the edge of the ischial bone to the back of the buttocks (mm) = 2.701 * hip circumference + 20.675.

Acknowledgment. This research is supported by “Special funds for the basic R&D undertakings by welfare research institutions” (522016Y-4680), National Key R&D Program of China (2017YFF0206602), National Science and Technology Basic Research (2013FY110200).

References

1. Tan, H.Z., Slivovsky, L.A., Pentland, A.: A sensing chair using pressure distribution sensors. IEEE/ASME Trans. Mechatron. **6**(3), 261–268 (2007)
2. Mutlu, B., Krause, A., Forlizzi, J., et a1.: Robust, low-cost, non-intrusive sensing and recognition of postures. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, pp. 149–158. ACM (2007)
3. Xu, L., Chen, G., Wang, J., et al.: A sensing cushion using simple pressure distribution sensors. In: 2012 IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 451–456. IEEE (2012)
4. Kamiya, K., Kudo, M., Nonaka, H., et a1.: Sitting posture analysis by pressure sensors. In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4. IEEE (2008)
5. Pynt, J., Higgs, J.: A History of Seating, 3000 BC to 2000 AD: Function Versus Aesthetics. Cambric Press, Amherst (2010)
6. Pynt, J., Higgs, J., Mackey, M.: Milestones in the evolution of lumbar spinal postural health in seating. Spine **27**(19), 2180–2189 (2002)
7. Mandal, A.C.: The seated man (Homo Sedens) the seated work position. Theory practice. Appl. Ergon. **12**(1), 19–26 (1981)

8. Mandal, A.C.: The correct height of school furniture. Hum.: J. Hum. Factors Ergon. Soc. **24**(3), 257–269 (1982)
9. Mandal, A.C.: The influence of furniture height on backpain. Behav. Inf. Technol. **6**(3), 347–352 (1987)
10. Noro, K., Naruse, T., Lueder, R., et al.: Application of Zen sitting principles to microscopic surgery seating. Appl. Ergon. **43**(2), 308–319 (2012)
11. Van, J.: Study: sitting up straight hurts your back. Chicago Tribune, (1 I) (2006)
12. Ma J., et a1.: Research of the comfort of the automotive seat. Shanghai Auto (12), 24–27 (2008)
13. Nuti, A.C.: Objective metric X subjective evaluation (2003)
14. Muraoka, T., Nakashima, N., Shimodaira, Y., et a1.: Subjective evaluation method for density unevenness of complicated Chinese characters in printing system. In: Proceedings of the IEEE International Symposium on Industrial Electronics, ISIE 1996, vol. 1, pp. 288–292. IEEE (1996)
15. Tatjana, P., Dubravka, M.: Subjective assessment of mastication as parameter for successful prosthetic therapy. Stomatoloski Glasnik Srbije **56**(4), 187 (2009)
16. Kushida, C.A., Chang, A., Gadkary, C., et al.: Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. Sleep Med. **2**(5), 389–396 (2001)
17. Vergara, M., Page, A.: Relationship comfort and back posture and mobility in sitting-posture. Appl. Ergon. **33**(1), 1–8 (2002)
18. Kang, T.E., Mark, A.F.: Development of a Simple Approach to Modify the
19. Hanel, S.E.: Measuring methods for comfort rating of seats and beds. Ind. Ergon. **20**, 163–177 (1996)
20. Wu, X., Rakheja, S., Boileau, P.E.: Distribution of human seat interface pressure on a soft automotive seat under vertical vibration. Int. J. Ind. Ergon. **24**, 545–557 (1998)
21. Groenesteijn, L., Ellegast, R.P., Keller, K., Krause, F.: Office task effects on comfort and body dynamics in five dynamic office chairs. Appl. Ergon. **43**, 320–328 (2012)
22. McCormi, E.J., Sanders, M.S.: Human Factors in Engineering and Design. McGraw-Hill Book Company, New York
23. Woodson, W.E.: Human Factors Design Handbook (1980)
24. Shackel, B., Chidsey, K.D., Shipley, P.: The assessment of chair comfort. Ergonomics **12**(2), 269–306 (1969)
25. Pankoke, S., Siefert, A.: Virtual simulation of static and dynamic seating comfort in the development process of automobiles and automotive seats: application of finite—element-occupant-model CASIMIR. In: Digital Human Modeling Conference (2007)

Effect of Mental Fatigue on Visual Selective Attention

Qianxiang Zhou[1,2], Jiaxuan Li[1,2], and Zhongqi Liu[1,2](✉)

[1] Key Laboratory for Biomechanics and Mechanobiology
of the Ministry of Education, School of Biological Science
and Medical Engineering, Beihang University, Beijing 100191, China
zqxg@buaa.edu.cn, flyhawk505@sina.com, lzq505@163.com
[2] Beijing Advanced Innovation Centre for Biomedical Engineering,
Beihang University, Beijing 102402, China

Abstract. To explore the effect of human mental fatigue on their visual selective attention, and carry out ergonomic design in monitoring operation system. Thirty-two men participated the experiment, 140 min digital 2-back task was used to simulate the monitoring work to induce mental fatigue, then simulation of automatic alarm based on 2 × 3 factors cue-target paradigm was used to measure the selective attention ability. The influence of mental fatigue on visual selective attention was explored by analyzing behavioral and event-related potentials data. The results showed that 140 min 2-back cognitive task induces mental fatigue successfully. Under the condition of mental fatigue, the invalid cue and the target stimulus in central cue-target paradigm had the strongest fatigue effect, while the effective cue and the interfere stimulus had the weakest. The subjects' performance of the selective attention task decreased and P300 latency was significantly prolonged and amplitude decreased significantly, mental fatigue has a negative effect on visual selective attention.

Keywords: Mental fatigue · Visual selective attention · 2-back · Performance · Electroencephalogram (EEG)

1 Introduction

With the rapid advancement of electronic informatization, the amount of information has increased dramatically, and the types of information that need to be processed by humans are richer and more numerous. The role of people in the system is no longer a simple operator, but tends to be more of a monitor and a decision maker. On the one hand, the complexity and variety of job tasks led to a significant increase in the operator's cognitive and mental workload; on the other hand, the operators were under tremendous mental stress while completing the task for a long time [1]. Therefore, operators were more likely to fall into fatigue due to long-term accumulation of mental workload and increased psychological stress. This mental fatigue was a limited state of the brain due to the completion of long repeated tasks [2].

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 236–245, 2019.
https://doi.org/10.1007/978-3-030-22507-0_19

The information processing capability of the brain is limited. Selective attention was to focus limited cognitive resources on important aspects of important things or things, which helps to improve information processing efficiency [3]. During the information operation, the operator needs to monitor the system status for a long time and make a correct response after sensing the system signal. This requires the operators had strong visual selective attention, maximize the operational ability, and improve information processing efficiency [4]. In the state of mental fatigue, the operators' attention ability and performance decreased, and the error rate increased. This situation might even lead to an accident [5]. Therefore, it is of great theoretical and practical significance to study the effects of mental fatigue on visual selective attention and to explore the quantitative changes of visual selective attention after mental fatigue.

It is difficult for people to concentrate when they are mentally fatigued. Existing research showed that mental fatigue induced by continuous cognitive operation has varying degrees of influence on selective attention [6–8]. For example: Zhang simulated 6 h driving task induced mental fatigue of the subject, and found that the individual noticed that the processing function was impaired, but the information processing speed was not damaged [6]; Mun et al. studied the effects of mental fatigue caused by moving 3D images on selective attention ability, and found that mental fatigue can impair spatial selective attention [7]; Faber et al. used a 120-min lateral inhibition task to induce mental fatigue, which also demonstrated that mental fatigue affects selective attention, and that the degree of fatigue increases, and the inhibitory ability of distraction stimulation decreases [8].

There are many studies on the effects of mental fatigue on selective attention, but the impact of its impact mechanism, especially on the visual selective attention ability of participating in surveillance operations, is unclear. At present, selective attention at home and abroad is based on basic research, and selective attention has not been combined with specific engineering applications [3, 4]. Therefore, the author will simulate the monitoring and alarm platform in the real scene, and use the memory refresh test (2-back task) to simulate the monitoring process to induce the mental fatigue process, in order to provide a theoretical basis for the operator's risk assessment under the condition of mental fatigue.

2 Method

2.1 Participants

A total of 32 subjects were recruited and randomly assigned to the normal and fatigue groups, each group of 16. All the subjects were male, aged 20 to 30 years old, bachelor degree or above, right hand, no red and green blindness, no difference in sensitivity to red and green color; normal cognitive ability, good sleep habits, no smoking, alcoholism Other bad habits, no irritating substances such as alcohol and caffeine were ingested within 4 h before the start of the test.

2.2 Experimental Tasks

The trial used the 140 min 2-back paradigm task, requiring participants to compare whether the current content is consistent with the target stimulus. And by constantly refreshing short-term memory, the nerve cells became fatigued, and the brain resources were quickly depleted, thus truly making the brain itself fatigued [9].

The visual selective attention test used a central clue (CC) to prompt the task, in which the cue clue appears at the center of the screen or does not appear, using a 2 × 3 design. And the independent variable A indicated the clue validity (effective clue (EC), invalid clue (IC), no clue (NC). Independent variable B indicated the stimulus color (red, green). The six situations of clue validity and stimulus color interaction corresponded to the automatic alarm platform hit (H), false alarm-0 (FA-0), false alarm-1 (FA-1), false alarm-2 (FA-2), missing report (M), no warning (N) 6 cases, and the parameter settings for the clue prompt task were shown in Table 1. The single subprocess (trial) of the clue prompt task consisted of 4 pictures, and a black block appeared inside the center ring of the leftmost picture. After a period of time, the position of the outer ring corresponding to the quadrant of the black block might appear as red or green targets. Red as the target represented a small probability event, and green as a disturbance represented a high probability event. The task interface was shown in Fig. 1. After the red target appeared, the subjects were asked to quickly press the "F" key with left index finger, and after the green interference occurred, to quickly press the "J" button with right index finger. Each trial contained 240 trials, of which the ratio of missing report, hitting and false alarm is 1:3:13. Subjects were asked to keep their eyes on the center of the word "Ten". Red as the target and green as the interference, the purpose was to pass the message "Red represents danger signal, and green represents safety signal" to the subject, so that the test interface and the operation of the subject matched the automatic alarm monitoring platform.

Table 1. Parameter setting for central cue-target paradigm

Position	Type	Color	Monitoring work	Press	Trials number
CC	EC	Red	H	F	36
		Green	FA-0	J	108
	IC	Red	FA-1	F	12
		Green	FA-2	J	36
NC	NC	Red	M	F	12
		Green	N	J	36

2.3 Procedure

The experiment was conducted as follow procedures:

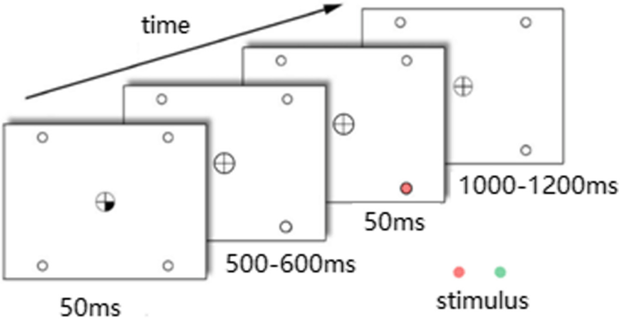


Fig. 1. The sketch map of task interface (Color figure online)

1. The subjects were asked to practice the experimental tasks to familiarize the trial operation and minimize the learning effect. The data in the practice session was not used for statistical analysis.
2. The two groups of subjects completed the 7 min clue prompt task (selective attention ability pretest).
3. Filled in the mental state subjective questionnaire (subjective questionnaire pretest).
4. The normal group and the fatigue group completed the 10-min 2-back task (2-back pretest).
5. The normal group rested for 120 min, while the fatigue group needed to continue to complete the 120 min 2-back task.
6. After the two groups of subjects completed the 10-min 2-back task (2-back post-test), the mental fatigue induction phase ended.
7. Two groups of subjects completed the subjective questionnaire of mental state (post-subject questionnaire) and completed the clue prompt task for about 48 min (selective attention ability post-test).

The entire test process took about 190 min. The electroencephalogram (EEG) signals were collected during the test. The scalp impedance was controlled to less than 5 kΩ and sampling rate was 1000 Hz.

3 Results

3.1 2-Back Mental Fatigue Induced Results

(1) Behavioral data

On the subjective evaluation data, there was no statistically significant difference in subjective fatigue before and after brain fatigue induction in the normal group. For the fatigue group, the subjective fatigue after brain fatigue induction was significantly higher than that before induction, and there were significant differences in thinking clarity, attention concentration, sleepiness, fatigue state and emotional state ($p < 0.05$). Therefore, 140 min of 2-back task significantly increased subjective fatigue.

Comparing the performance data of the 2-group pre-test and post-test tasks in the normal group operation, it was found that the difference in response time, stability, error rate and miss rate of the pre-test and post-test tasks was not statistically significant. Comparing the performance data of the 2-back pre-test and post-test tasks of the fatigue group, it was found that the response time of the post-test task decreased, but it was not statistically significant ($t = 1.219$, $p = 0.233$), and the stability did not change significantly ($t = -0.029$, $p = 0.977$), the error rate increased significantly ($t = -2.276$, $p = 0.032$), and the missing rate increased slightly, but it was not statistically significant ($t = -1.290$, $p = 0.208$). The above analysis validated the effectiveness of the 140 min 2-back task-induced mental fatigue from the perspective of behavioral data.

(2) **EEG data**

P300 components were extracted from EEG data to analyze the correct response of the subjects to the stimulus corresponding to the operation' EEG fluctuations. The magnitude of P300 reflected the amount of mental resources invested in performing cognitive tasks. The length of the incubation period reflected the speed of cognitive processing [10].

The P300 amplitude and latency of the 2-back pre-test and post-test tasks were analyzed statistically in the normal group and the fatigue group. It was found that there was no significant change in the P300 amplitude and latency of the normal group before and after the mental fatigue induced task. Therefore, from the perspective of EEG data, it was proved that the 20 min 2-back task did not cause mental fatigue. Before and after the brain fatigue induced task, the P300 amplitude of the fatigue group did not change significantly, and the latency was significantly prolonged ($p < 0.05$).

The 140-min 2-back operation in the fatigue group was divided into 14 segments on average (i.e., pre segment, 1–12 segment, post segment, each segment 10 min). The changes of P300 amplitude and latency of each segment with time were shown in Figs. 2 and 3. Fz, Cz and Pz electrodes were selected to represent three brain regions: the middle frontal region, the vertex region and the apex region respectively. In the first 50 min of the 2-back task, P300 amplitude and latency showed an upward trend, indicating that the brain resources of the participants were increasing, but the cognitive process of the stimulus signal became slower, and the time required for information perception and processing was prolonged. It showed that the fatigue of the subjects is deepening. At the 60th minute of the 2-back task, the amplitude of P300 suddenly dropped, indicated that the subject did not mobilize too much brain resources, and the effort on information processing decreased. At the same time, the latency of P300 decreased slightly, indicating that the processing speed of stimulus information was slightly accelerated, indicating that the operation of the subject became sloppy at this time. 80 min after the 2-back task (paragraph 6-post), the fluctuation range of P300 amplitude was large, which was believed to be caused by the subject's subjective self-motivation and adjustment. The latent period of P300 presented an upward trend, indicating that the subject was still in a state of mental fatigue during this period. Based on the above analysis, it was further proved from the perspective of EEG that the 140 min 2-back task successfully induced mental fatigue.

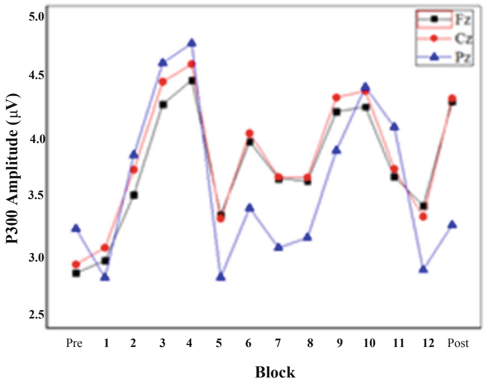


Fig. 2. Each segment's P300 amplitude of fatigue group (Color figure online)

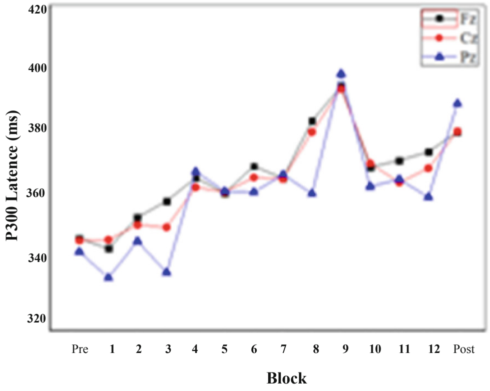


Fig. 3. Each segment's P300 latency of fatigue group (Color figure online)

3.2 Visual Selective Attention Ability Test Results

(1) Behavioral data

Table 2 was the descriptive statistics of the performance data of operation clues and post-test tasks in the normal group (NG) and the fatigue group (FG). Compared with the normal group, the error rate (ER) and the missing rate (MR) of the fatigue group increased. The reaction time (RT) increased, and the operation performance deteriorated.

Statistical analysis of the normal group and fatigue group of the performance of the operation clues hinted after the test task data. In the reaction time, the main effect of the grouping factor was significant (F = 4.446, p = 0.044), and the interaction effect of color * group was significant (F = 7.617, p = 0.010); Simple effect analysis was carried out, and it was found that the effect of grouping factors was affected by color factors. When only green stimuli appeared, there were statistical differences in reaction

Table 2. Performance data on the normal group and the fatigue group after post test task (x ± s)

		EC		IC		NC	
		Red	Green	Red	Green	Red	Green
NG	RT (ms)	449.200	337.667	469.333	428.429	493.400	410.733
		(49.124)	(40.739)	(57.568)	(66.232)	(53.341)	(49.117)
	ER	0.094	0.005	0.047	0.006	0.063	0.005
		(0.064)	(0.004)	(0.037)	(0.005)	(0.046)	(0.005)
	MR	0.014	0.012	0.014	0.020	0.011	0.034
		(0.015)	(0.009)	(0.015)	(0.023)	(0.009)	(0.034)
FG	RT (ms)	474.929	389.286	494.571	381.000	524.000	486.929
		(67.722)	(63.585)	(70.099)	(43.619)	(64.649)	(73.921)
	ER	0.100	0.013	0.083	0.019	0.075	0.011
		(0.100)	(0.012)	(0.076)	(0.024)	(0.079)	(0.015)
	MR	0.049	0.045	0.045	0.052	0.060	0.092
		(0.059)	(0.053)	(0.057)	(0.063)	(0.072)	(0.095)

time between normal group and fatigue group (F = 0.138, p = 0.008). At this time, the effect of mental fatigue on wireless clue was greater than the effective clue, and the degree of influence on effective clues was similar to that of invalid clues. In terms of error rate, the main effect of grouping factors was not significant (F = 1.000, p = 0.326), and there was no significant interaction between grouping factors and other factors. In the miss rate, the main effect of the grouping factor was significant (F = 6.107, p = 0.020), and there was no significant interaction between the grouping factor and other factors.

Further analysis of the main effects and interaction effects of intra-group factors (color, validity), was made that the main effect of color factors was significant in response time (F = 141.227, p = 0.000, red > green). Effectiveness factors of the main effect was significant (F = 110.777, p = 0.000). In terms of error rate, the main effect of the color factor was significant (F = 40.620, p = 0.000, red > green). When and only when the red stimulus appears, there was a significant difference in the error rate of the subjects under different clues. The error rate of the subject under the valid clue was significantly larger than the invalid clue (F = 0.032, p = 0.002) and Wireless clue (F = 0.028, p = 0.021). In the miss rate, if and only if the wireless clue prompts, the miss rate for green stimulus was significantly greater than the red stimulus (F = −0.028, p = 0.001). If and only if the green stimulus appears, the miss rate of the participant under the wireless clue prompt was significantly larger than the effective clue (F = −0.035, p = 0.001) and the invalid clue (F = −0.027, p = 0.004).

Therefore, under the clue prompt, the subject selectively paid attention to the large probability event, and the operation of the small probability event was easy to make mistakes. Under the wireless clue prompt, the subject selectively paid attention to the small probability event, and the operation of the large probability event was easy to be missed. Mental fatigue caused the subject's performance to deteriorate. The effect of mental fatigue on wireless clue was greater than that of clues, and the degree of influence on effective cues was similar to that of invalid cues. The effect of mental fatigue on red stimuli was greater than that of green stimuli.

(2) **EEG data**

Statistical analysis of the normal group and the fatigue group operating clues prompted the post-test task P300 amplitude, latency, descriptive statistical results were shown in Table 3. Compared with the normal group, the latency of P300 in the fatigue group was significantly prolonged ($p < 0.05$). The amplitude of the P300 was significantly reduced ($p < 0.05$).

Table 3. P300 amplitude and latency of the cue-target paradigm in the normal group and the fatigue group (x ± s)

Factors		EC		IC		NC	
		Red	Green	Red	Green	Red	Green
NG	Amplitude/μV	3.93	1.89	3.99	2.14	3.46	1.46
		(1.48)	(1.17)	(1.74)	(1.49)	(1.62)	(0.80)
	Latence/ms	377.43	386.63	389.98	403.21	391.00	403.41
		(13.31)	(15.47)	(9.85)	(11.85)	(11.78)	(12.36)
FG	Amplitude/μV	3.33	1.30	3.11	1.23	2.23	1.08
		(1.49)	(0.72)	(1.96)	(0.65)	(1.49)	(0.63)
	Latence/ms	390.59	388.28	397.06	413.84	420.82	406.82
		(11.71)	(17.30)	(6.09)	(10.13)	(9.88)	(10.33)

For the P300 amplitude, the main effect of the grouping was significant ($F = 17.314$, $p = 0.000$), and the main effect of the color was significant ($F = 194.088$, $p = 0.000$). The main effect of validity was significant ($F = 12.167$, $p = 0.000$), and the interaction effects were not significant. For the P300 latency, the main effect of the grouping was significant ($F = 56.894$, $p = 0.000$), and the color * validity * grouping interaction was significant ($F = 3.348$, $p = 0.000$). Further analysis of the interaction effect found that the fatigue effect of red stimulation and green stimulation was significant ($F = 96.570$, 5.840, $p = 0.000$, 0.022); effective clues, invalid clues, fatigue effects of wireless cables was significantly ($F = 7.190$, 17.270, 96.530, $p = 0.012$, 0.000, 0.000).

The results of P300 analysis showed that the fatigue effect of invalid clues and red stimuli is the strongest, and the fatigue effect of effective clues and green stimuli was the weakest, and indicated that there are more brain resources required for invalid cues and small probability events, and less brain resources required for effective cues and high probability events.

4 Discussion

The 140-min digital 2-back task successfully induced mental fatigue. By analyzing the behavior and the P300 data of the fatigue group, it was found that the subjective fatigue of the fatigue group was significantly higher than that of the normal group, and the reaction time and stability showed strong fluctuations. Error rate and omission rate showed an upward trend. On the EEG data, the P300 amplitude of the fatigue group showed strong volatility, and the P300 latency period showed an upward trend. It was found that the subject reach the mental fatigue state after 60 min. With the extension of the task time, the degree of mental fatigue was constantly deepening. Consistent with previous studies, Horat et al. also found that the P300 latency is prolonged as mental fatigue deepened [11].

Mental fatigue affects visual selective attention ability, and the results of performance data showed that mental fatigue leads to poor performance of subjects, which has a greater impact on non-cue than on cue, and a similar impact on effective cue and invalid cue. The results of P300 showed that the fatigue effect of invalid clues is the strongest, while that of effective clues was the weakest. There were some differences between this conclusion and the conclusion of performance analysis. The reason was that human behavior is not only affected by the cerebral cortex, but also involves a series of executive control links. Performance data cannot accurately reflected the real state of mind, and Event Related Potentials (Event - Related Potentials, ERPs) as a cognitive characteristics study of the characteristics of the most direct, strongest time resolution could be better objectively reflect the state of cognitive and mental load, is an effective indicator of cognitive and mental fatigue analysis. The fatigue effect of red stimulus was the strongest, while the fatigue effect of green stimulus was the weakest, indicating that more mental resources are needed for invalid clues and low-probability events, while less mental resources are needed for effective clues and high-probability events. This conclusion was consistent with the conclusion of performance analysis. Green stimulation was a high probability event. In comparison, the process of the high probability event was more automated during the test. Therefore, the brain resources needed to respond correctly to green stimuli are few. The red stimulus was a small probability event. To respond correctly to the red stimulus, more brain resources needed to be called to identify and judge the stimulus. In consequence, the response to green stimuli was faster and more accurate, and the response to red stimuli was slower and more error-prone.

5 Conclusion

Brain power fatigue was induced by continuous operation of the digital 2-back task for 140 min, and the selective attention ability was measured by the central cue paradigm to explore the effect of brain fatigue on visual selective attention. Conclusions were drawn as follows:

By analyzing behavioral data and EEG data, it was proved that the 140 min 2-back task successfully induces mental fatigue. Therefore, when the task is completed for a long time, the mental fatigue will be induced, which will affect the performance of the operation. The main solution at this stage was the shift and rest of the personnel. In the future, the adaptive assignment task system can be applied to dynamically assign tasks according to the personnel load status, thereby improving man-machine compatibility.

The negative impact of mental fatigue on the ability of visual selective attention leads to poor performance of the test, prolonged latency and reduced amplitude of P300; the fatigue effect of invalid clues and target stimulation is the strongest, and effective clues and interference stimulation is the weakest. Therefore, the number of invalid leads should be appropriately reduced when designing the monitoring operating system.

This article is only a preliminary study of mental fatigue. In the future, other components of ERPs need to be extracted, combined with ECG, eye movement and other physiological indicators, to further explore the indicators of mental fatigue and selective attention.

References

1. Zhao, S.Q., Jiang, Y., Song, Y.: Research on design and application of command and training system based on integrated command platform. Mil. Oper. Res. Syst. Eng. **3**, 35–39 (2011)
2. Xiao, Y., Ma, F., Li, Y.X.Y., et al.: Sustained attention is associated with error processing impairment: evidence from mental fatigue study in four-choice reaction time task. PLoS One **10**(3), e0117837 (2015)
3. Couper, U.J.W., Mangun, G.R.: Signal enhancement and suppression during visual-spatial selective attention. Brain Res. **1359**, 155–177 (2010)
4. Xie, S.Y., Wang, L.N., Zhang, B., et al.: On spatial selective attention pattern and its application to monitoring of pilot's brain function state. J. Northwest. Polytechnical Univ. **32** (2), 268–272 (2014)
5. Sun, R.S., Ma, G.F., Yuan, L.P.: Analysis of risk of controller fatigue based on characteristics of speech reaction time. China Saf. Sci. J. **26**(12), 7–12 (2016)
6. Zhang, Y.: The effect of mental fatigue upon attention characteristics. Fourth Military Medical University, Xi'an (2009)
7. Mun, S., Kim, E.S., Park, M.C.: Effect of mental fatigue caused by mobile 3D viewing on selective attention: an ERP study. Int. J. Psychophysiol. **94**(3), 373–381 (2014)
8. Faber, L.G., Maurits, N.M., Lorist, M.M.: Mental fatigue affects visual selective attention. PLoS One **7**(10), e48073 (2012)
9. Tanaka, M., Shigihar, A.Y., Ishii, A., et al.: Effect of mental fatigue on the central nervous system: an electroencephalography study. Behav. Brain Funct. **8**(2), 262–270 (2012)
10. Murata, A., Uetake, A., Takasawa, Y.: Evaluation of mental fatigue using feature parameter extracted from event-related potential. Int. J. Ind. Ergon. **35**(8), 61–77 (2005)
11. Horat, S.K., Herrman, F.R., Favr, E.G., et al.: Assessment of mental workload: a new electrophysiological method based on intra-block averaging of ERP amplitudes. Neuropsychologia **82**, 11–17 (2016)

Cognitive Psychology in Aviation and Space

Analysis of Key Cognitive Factors in Space Teleoperation Task

Junpeng Guo(✉), Yuqing Liu, Xiangjie Kong, Shihua Zhou, and Jin Yang

National Key Laboratory of Human Factors Engineering,
China Astronaut Research and Training Center, Beijing, China
dragonguo@126.com

Abstract. Teleoperation is always a challenging task due to the lack of sense of immediacy and insufficient information for operation. Robotic arm operation in space used for transporting astronauts during the extra-vehicle activities or docking spaceships is one of such tasks. In this study, we proposed a simplified hierarchical model that could describe the space teleoperation process. And according to this model, the cognitive factors that might influence the teleoperation task were analyzed. In order to verify the hierarchical model and the effects of the factors, an experiment was conducted via a computer simulation platform.

Keywords: Space teleoperation · Hierarchical model · Cognitive factors

1 Introduction

Teleoperation is always a challenging task due to the lack of sense of immediacy and insufficient information for operation. Robotic arm operation in space used for transporting astronauts during the extra-vehicle activities or docking spaceships is one of such tasks. Safe and efficient control of the robotic arm is heavily dependent on the spatial skills of the operator [1], so it is important to make it clear the key cognitive factors in the space teleopertation task, especially for improving the astronauts' training efficiency.

According to NASA's related report [2], during the robotic arm operation task in space, astronauts usually have to make decisions only relying on the visual feedback from the cameras mounted on the robotic arm and at various locations on the space station exterior to learn about the spatial relationship of the arm with the surrounding structure. However, there are usually only three camera viewpoints available at any moment [3], and the cameras are not always placed at the optimal locations for astronauts to observe the clearance from the structure. Addition to these, astronauts also have to memorize the location of these cameras and switch between them, which undoubtedly increases the operators' mental workload during the operation. To avoid the danger of colliding with structure or singularities during the operation, the operators need to confirm they handle the hand controllers correctly before they give that command, and the procedures may also require a second operator to provide additional monitoring of the scene. At the same time, the movements of the robotic arm are made

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 249–258, 2019.
https://doi.org/10.1007/978-3-030-22507-0_20

very slow and after the operators have established situation awareness of the spatial relationship [4] to guarantee safety.

Under this circumstance, the preflight robotic arm training on the ground becomes very important for the operation success in the orbit. Astronauts must experience sufficient training first before they execute a real task so that they can be skilled enough to operate successfully. The robotic arm training for the astronauts in NASA began with a generic robot arm simulation, which having a 6 degrees-of-freedom robot arm and different camera views available as well as two hand controllers [5]. Trainees needed to learn how to visualize the clearance and choose the ideal cameras during the operation process. They also needed to learn how to make the right control commands via the hand controllers to avoid collisions or singularities. During the training process, the operator's performance was usually evaluated by a robotics instructor and an instructor astronaut according to standard criteria covering all aspects of operations [6]. The criteria could include spatial/visual perception, situational awareness and appropriate input of the controllers. However, the performance scores given by instructors could be subjective sometimes, so we also need some objective criteria, especially relating to the operation process and operation efficiency.

Usually, the training process took much time. In order to improve the training efficiency and make the training course more customized, the key cognitive factors are needed to be identified first. Previous studies showed that individual spatial ability might be related with the operation performance [7]. In this study, we mainly focused on analyzing the key cognitive factors influencing the space teleoperation task for training, which was important to cultivate and maintain the operation skills of operator. To achieve this, we needed to observe the operation training procedure. But on the ground it was not easy to reproduce the working environment physically, so we built up a simulated space robotic arm operation platform by means of computer simulation, which could provide various operating conditions similar to a real space robotic arm aboard the space station and the operator could use it via the hand controllers. With the operating experience in this simulated environment, we constructed a hierarchical model that described the space robotic arm operation task. And then the key cognitive factors that might affect the task training process were proposed according to the hierarchical model. Lastly an experiment in this simulated operation platform was conducted to testify the effect of these cognitive factors.

2 The Hierarchical Model for the Robotic Arm Operation Task

In order to analyse the key cognitive factors influencing the space teleoperation task training, we should firstly learn about the procedure in the robotic arm operation. From the review of the in-orbit robotic arm operation activities, the kinds of operation tasks could be various [8, 9]. For example, the robotic arm could be used for satellite deployment and retrieval, docking with the space station and transporting astronauts during extra-vehicle-activities and so on. Although the contents of the operation might be varied in different tasks, they also had some potential characteristic in common. During each operation task, the operator needed to observe the status (e.g. position and attitude) of the robotic arm firstly, so that the operator could be aware of the spatial relationship between the arm and the exterior of the space station. This was the fundamental step of the whole subsequent operation steps, and we named this step of the operation as “Evaluate the present situation”; and then, immediately after this step the operator would be able to establish the awareness of the present situation, which made him/her possible to foresee the integrated route for the movement of the robotic arm and plan for the operation commands for next few steps, so we classified this step of the operation as “Plan for the following operation commands”; and finally, the operator needed to transform the operation plan into real control commands via hand controllers, and confirmed whether the outcome of the control command met his/her expectations, and we defined this final step as the “Execute the commands and confirm the realtime outcome”. The three steps described above could be regarded as a small operation unit of the whole robotic arm operation. During the whole operation procedure, this operation unit would be repeated periodically until the operation ended in success (Fig. 1).

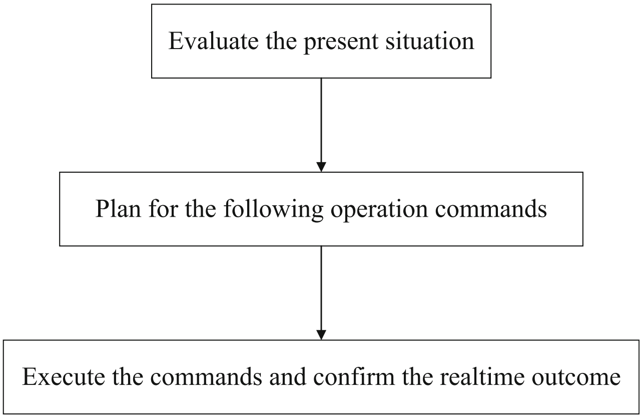

Evaluate the present situation
Plan for the following operation commands
Execute the commands and confirm the realtime outcome

Fig. 1. The illustration of the hierarchical model that described the process of the space teleoperation task investigated in this study.

3 The Cognitive Factor Analysis in Space Teleoperation Task Training

From the hierarchical model described above, we could extract the key factors that influenced the operation process.

In the first step, namely the step “Evaluate the present situation”, the operator would observe the status of the robotic arm and its spatial relationship with the exterior of the surrounding structure via different cameras mounted at different locations. These cameras could be located at the exterior of the space station as well as at the end of the robotic arm, and the operator could view the images from several different cameras at the same time. Due to the number of the operation monitors in the robotic arm workstation, usually there were only three or four images available at any moment, so this would require a camera selection process when evaluating the present situation [10]. The operator must select the most suitable three or four camera images from the whole cameras for observing the position and attitude of the robotic arm and determining the clearance distances.

The cameras could be divided into two categories. The first kind were the cameras located at the exterior of the space station or the base of the robotic arm. The images from this kind of the camera were presented from a fixed view and would not change as the robotic arm moved. These cameras could provide stable view images for the operators, which could give operators an exocentric view frame of the robotic arm movement [11]. The second kind of the cameras were the ones located at the end of the robotic arm. The images from this kind of the camera could provide a view attached to the end of the arm, from which the operators would obtain a sense of egocentric frame understanding of the environment. These two different categories of the cameras could result in two different perception mode. The first perception mode was related to the first kind of cameras. In this mode, the operator would perceive the status of the robotic arm in an exocentric way, which might help the operator form an overview of the situation. The second perception mode was related to the second kind of cameras, and in this mode, the operator perceived the situation of the robotic arm in an egocentric way, which could provide the operator a more specific view about the situation of the surroundings around the arm. As these two different perception modes might influence the way operators observe the robotic arm status, so the perception mode could be one of the factors influencing the teleoperation task.

In the second step, the operator needed to plan for the following operation commands. During this stage, the operator should form a plan for how to give the appropriate following operation commands based on the evaluation results from the first step. For example, if the robotic arm was evaluated to be pitched up too much in the previous step, then the operator needed to know how to adjust its attitude via proper operation commands in this step. In order to achieve this, the operator needed to take advantage of the mental imagery to make a rehearsal of the robotic arm's trajectory as it was hard to predict the movement of the arm accurately and in time through other methods. As a result, the operator should be able to image what the position or attitude of the robotic arm would be after certain commands as well as the perspective of the arm from other views that was not presented in the current monitors. This mental imagery transformation process was similar with the two common-used individual spatial ability factors, namely spatial visualization and perspective-taking abilities [12]. These two kinds of spatial abilities might be the two related cognitive factors. Spatial visualization ability was also usually presented by mental rotation ability. It was referred to the ability to mentally manipulate an array of objects. The manipulation was in a fixed egocentric reference frame. Perspective-taking ability described the ability to imagine how an object or scene looked from perspectives different to the observer's current view. It demanded a transformation process in the egocentric reference frame while the world coordinate frame was fixed. These two abilities could be regarded as logically equivalent, the only critical difference was in the coordinate frame which was manipulated to obtain the final view. Previous studies showed that although the performance was also highly correlated, a measurable distinction between spatial visualization and perspective-taking ability was found [13]. So we needed to consider these two cognitive factors in separated ways.

In the third step, the operator's task was to execute the commands and confirm the realtime outcome. When executing the commands during this procedure, there existed two different possible types of control modes. The main difference between these two modes was the coordinate system used. This could effect the methods of adjusting the position and attitude of the arm. In the first mode, the origin of the control coordinate system was at the end of the robotic arm and the norm's direction changed accordingly while the attitude of the arm varied, we named this mode as the "end-control-mode". Under this circumstance, the control direction of the position and attitude was coupled. For example, in this control mode, the upward direction would changed for 90° after pitching the same extent. So it would require the operator to transform the mental representation of the control direction constantly with the movement of the arm, which might increase the mental workload, but on the other side, could give the direct commands relating to the arm's current status, and it might be helpful for improving the situation awareness of the operator. In the other control mode, the origin of the control coordinate system was located at some point at the exterior of the space station, e.g. it could be at the base of the robotic arm and so it would be stable during the whole process, and we named this mode as the "global-control-mode". And also the norm of the coordinate system would keep stable. This could provide a constant reference frame for the robotic arm's operating and none of the control directions was changeable no matter how the movement rotated. In this situation, the control direction of the position and attitude was decoupled. This could be helpful when the operator received the images from the exocentric cameras, but might also confuse the operator when the images was presented in an egocentric camera. Although in the second step the strategy used for operating was formed, the control mode was also needed to be taken into consideration to obtain the final commands to be executed. So we considered that the control mode could influence the space teleoperation task operation.

From the analysis above, we concluded that there might exist three different kinds of factors that influenced the space teleoperation task, namely perception mode, individual spatial ability and control mode. These three steps composed an operation unit during a complete teleoperation task and this unit was repeated periodically in the task. The task training process was also consisted of large amount of these repeated unit, so we could regard it that these factors drawn from the hierarchical model would likewise influence the space teleoperation task training process.

It was important and useful to identify these factors in theory as this was the first step for the continued study. And to go further, we conducted an experiment to testify the influence of the individual spatial ability factor. In this experiment, we built up a simulated space teleoperation platform that could be controlled via two hand controllers.

4 The Experiment Validation

4.1 Materials and Methods

Participants. Twenty-four adults (mean age = 26.3, SD = 2.75, ranging from 23 to 31) with college-level education participated the experiment. None of these participants had conducted the space teleoperation task in simulated or physics environment before the experiment. The study was approved by the IRB and all participants signed the informed consent prior to the experiment.

Measurement of Spatial Ability. Because two factors of the individual spatial ability, mental rotation and perspective-taking, were concerned by us mostly in this experiment, we measured both in their 2D and 3D versions. The 3D Mental Rotation Ability (MRA) was measured by Cube Comparison Test (CCT) on computer, and the 2D MRA was also measured by Card Rotation Test (CRT) through computer. The 3D Perspective-taking Ability (PTA) was measured by the paradigm developed by Guay [14] on computer, and the 2D PTA was also measured by the paradigm developed by Kozhevnikov and Hegarty [15] using specially developed software. The specific parameters used in the spatial ability tests were shown in Table 1.

Table 1. Parameters setting in the four spatial ability tests

Parameters	Trial numbers	Time limitation per trial	Test platform	Performance indicator
3D mental rotation	48	25 s	Computer software	Latency/percent correct
2D mental rotation	48	25 s	Computer software	Latency/percent correct
3D perspective-taking	24	40 s	Computer software	Latency/percent correct
2D perspective-taking	24	25 s	Computer software	Latency/percent correct

Simulated Space Teleoperation Task Platform. In this study, we developed a space teleoperation task platform via computer simulation. The participants could use the hand controllers to manipulate the simulated robotic arm in the in-orbit background. The difficulty of the tasks could be set differently with various parameter settings through the platform. These main parameters included the position of the target, the initial status of the robotic arm and the camera selection at the beginning and the two control modes as described above and so on. The user interface of the simulated platform was shown in Fig. 2. To evaluate the performance of the participants, the time consumed during the whole operation task and whether each task was finished successfully was recorded.

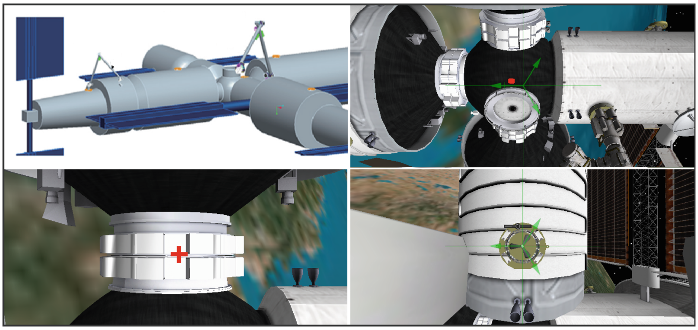


Fig. 2. The illustration of the overview [16] (the upper left) and the user interface of the simulated robotic arm platform. This illustration showed the camera views that available by the participant during one trial. This was the final stage of one trial and the red cross/star marked the target point that needed to approach to. (Color figure online)

Experiment Procedure. The whole experiment was conducted in two periods. In the first period, we measured the spatial ability of all the participants and calculated the participants' scores in the spatial ability tests; and then in the second period, after identifying all the participants' spatial ability scores were valid, the participants were arranged to take the simulated robotic arm operation experiment.

To be specific, during the first period, participants' spatial ability in 2D & 3D MRA and PTA were tested. Their scores in each test were obtained using the performance indicators as shown in Table 1. Then in the second period, there were twelve formal trials and all the participants needed to operate the simulated robotic arm twelve times after six practices. During the practice stage, the participant could ask the experimenter questions about how to use the hand controllers, but no help about the manipulation strategy were provided.

As in this experiment we mainly focused on the influence of individual spatial abilities on the task, we provided the same parameters about the initial status of the robotic arm as well as the camera selection at the beginning and the control mode for all the participants. To be specific, in all the twelve tasks, we provided the participants four camera views, two of which provided the exocentric perception mode and the rest two provided the egocentric perception mode. And the control mode was set to be the "global-control-mode". The task in each trial was to transform the simulated astronaut on the end of the effector to the target point. The main differences among the twelve formal tasks were the locations of the targets. The participants needed to transform the simulated robotic arm from the beginning position towards the target in each task. And in order to have enough complexity for each task, the locations were set to be accessible only after a combination of operations including pitch, yaw as well as roll. In order to avoid the learning effect, the sequence of the tasks were randomized for each participants.

4.2 Results

Statistical analysis was conducted using SPSS. We first observed the participants' performance in spatial ability tests. As both the latency and percent correct indicator could reflect the participants' performance, we use the ratio of latency and percent correct as the performance indicator to take both indicators into consideration at the same time, and marked as spatial ability synthetical indicator. The higher value of this indicator meant the better performance of the participant. Then we investigated the participants' performance in simulated robotic arm operation. As we provided enough practices before formal trial and did not set the time limitation to finish each trial, almost all the trials of every participant's was ended in success (only 2 of the all 288 trials were failed and were due to the inattentive judgement during the final stage). So we took the average time consumed in the operation task as the indicator that reflected the participants' performance in the task.

After the indicators was built up, the correlation between participants' spatial ability and simulated space teleoperation performance was analyzed. The correlation coefficients were shown in Table 2. From these results we could see that, except the result of the 2D mental rotation test, the other three of all the four spatial ability tests were correlated significantly.

Table 2. Pearson correlation coefficients for spatial ability test scores and simulated robotic arm operation task performance

Test type	3D mental rotation	2D mental rotation	3D perspective-taking	2D perspective-taking
Averaged time consumed in simulated robotic arm operation	−0.238**	−0.413	−0.179**	−0.259*

$^{**}p < 0.01$; $^{*}p < 0.05$.

4.3 Discussion

From the result described above, we could conclude that the mental rotation and perspective-taking ability of the individual spatial abilities were correlated with the performance in the simulated robotic arm operation. This result was consistent with the analysis we conducted above using the hierarchical model. Although the result of the 2D mental rotation was not correlated with the performance in robotic arm operation significantly, we thought this might be due to the robotic arm operation was mainly in the three-dimensional environment, and especially the transformation of the arm movement's mental representation seldom happened just in a two-dimensional world. However, the perspective-taking process in 2D and 3D were essentially related with each other closely, and the result also showed that both these two components were correlated with the operation performance.

5 Conclusions

Space teleoperation task is challenging for astronauts. It is important to identify the factors that may influence the operators' performance. In this study, we proposed a different way to explore the potential cognitive factors that might effect the robotic arm operation. Firstly, a hierarchical model was proposed and the operation process was divided and described by this model. Then according to the analysis of this model, we proposed some factors that might influence the teleoperation task training process. And finally we conducted an experiment via the computer-simulated method to identify one of the factors we proposed in theory, and the result was consistent with our previous analysis generally and the detailed discussion was given.

Except for the individual spatial ability factor, the remaining two other factors were still needed to be tested and verified in future studies. And in this study, we just focused on the space operation without time delay or the object on the end effector of the space robotic arm was cooperative, so the hierarchical model could was simplified at some extent. The method to establish a complete model with the time delay and cooperative object is also needed to be studied in the future.

Acknowledgments. This study was supported by the National Natural Science Foundation of China (No. 61772481). Author Yuqing Liu and Xiangjie Kong were supported by the foundation of National Key Laboratory of Human Factors Engineering (No. SYFD160091802). Author Jin Yang was supported by the foundation of National Key Laboratory of Human Factors Engineering (No. SYFD160051807).

References

1. Chen, J.Y., Haas, E.C., Barnes, M.J.: Human performance issues and user interface design for teleoperated robots. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **37**(6), 1231–1245 (2007)
2. Goza, S.M., Ambrose, R.O., Diftler, M.A., Spain, I.M.: Telepresence control of the NASA/DARPA robonaut on a mobility platform. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 623–629. ACM, April 2004
3. Pontillo, T.M.: Spatial ability and handedness as potential predictors of space teleoperation performance (Doctoral dissertation, Massachusetts Institute of Technology) (2010)
4. Allard, P.: A Virtual Environment for Training Space Station Teleoperators, pp. 1–85. McGill University (1997)
5. Liu, A.M., Oman, C.M., Galvan, R., Natapoff, A.: Predicting space telerobotic operator training performance from human spatial ability assessment. Acta Astronaut. **92**(1), 38–47 (2013)
6. Liu, A.M., Oman, C.M., Natapoff, A.: Efficient Individualized Teleoperation Training via Spatial Ability Assessment. NSBRI Final Report (2008)
7. Lamb, P., Owen, D.: Human performance in space telerobotic manipulation. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology, pp. 31–37. ACM, November 2005
8. Menchaca Brandan, M.A.: Influence of spatial orientation and spatial visualization abilities on space teleoperation performance (Doctoral dissertation, Massachusetts Institute of Technology) (2007)

9. Hoppenot, P., Rybarczyk, Y., Mestre, D., Colle, E.: Space perception for remote controlled tasks. In: IMACS 2005, p. na (2005)
10. Lapointe, J.F., Dupuis, E., Hartman, L., Gillett, R.: An analysis of low-earth orbit space operations (2002)
11. Lathan, C.E., Tracey, M.: The effects of operator spatial perception and sensory feedback on human-robot teleoperation performance. Presence: Teleoper. Virtual Environ. **11**(4), 368–377 (2002)
12. Menchaca-Brandan, M.A., Liu, A.M., Oman, C.M., Natapoff, A.: Influence of perspective-taking and mental rotation abilities in space teleoperation. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, pp. 271–278. ACM, March 2007
13. Hegarty, M., Waller, D.: A dissociation between mental rotation and perspective-taking spatial abilities. Intelligence **32**(2), 175–191 (2004)
14. Guay, R., Mc Daniels, E.: The visualization of viewpoints. The Purdue Research Foundation, West Lafayette (as modified by Lippa, I., Hegarty, M., Montello, D.R. (2002)) (1976)
15. Kozhevnikov, M., Hegarty, M.: A dissociation between object-manipulation spatial ability and spatial orientation ability. Mem. Cogn. **29**, 745–756 (2001)
16. Liu, H., Jiang, Z., Liu, Y.: Review of space manipulator technology. Manned Spaceflight **21** (5), 435–443 (2015)

Enhancing Aviation Simulator de-Briefs Through the Integration of Student Eye Tracking: The Instructor's Perspective

Julius Jakubowski[1](✉) and Wen-Chin Li[2]

[1] Royal Air Force Brize Norton, Carterton, UK
julesjakubowski@gmail.com
[2] Safety and Accident Investigation Centre, Cranfield University, Cranfield, UK

Abstract. Many research papers investigate the benefits to a pilot of eye tracking (ET) in training [1, 2]. None however, explore the efficacy of ET training from an instructor's perspective, specifically relating to trainee checking, scanning and monitoring. Using a questionnaire of both numerical rating and open answers, 19 experienced A330 instructors answered questions on simulator profiles that they observed on an operational simulator A330 De-Brief Facility (DBF). The profiles were played first without and then with ET data run in parallel. This research interrogated the ease with which instructors were able to interpret the data, investigated their ability to spot checking and scanning errors and challenged their pre-conceptions over how many errors they were able to identify. Results showed that considerable numbers of checking and scanning errors were missed and that with the augmentation of ET, error identification significantly increases. It was shown that instructors were not as error aware as they thought; moreover, prior to the addition of ET, instructors' focus was not on scanning and checking errors, despite the error presence.

Keywords: Attention distribution · Eye tracking · Scan patterns · Training evaluation · Visual behaviour

1 Introduction

Aviation Full Flight Simulators can record and playback almost every element of a simulation to aid instruction. What cannot be monitored is where a pilot is looking. From their very first flight, all pilots will, at some time, conduct insufficient monitoring and instrument checks. An instructor's inability to identify a lack of checking can lead to a normalization of deviance towards bad practice. As a pilot looks forwards in a simulator their face generally cannot be seen and their scanning accuracy cannot be monitored by an instructor if no subsequent failure or omission occurs. At present there are no aviation simulators that use ET technology to aid instruction despite countless articles espousing its benefits. Also notable is no research investigates how instructors respond to this technology. What use is all the data in the world, if it cannot be effectively interpreted? This paper discusses the importance of checking and scanning and then, through taking the instructor's perspective, investigates whether ET of student pilots could be used to assist simulator instructors in identifying and enhancing

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 259–270, 2019.
https://doi.org/10.1007/978-3-030-22507-0_21

their awareness of student errors. It identifies whether instructors maintain inaccurate preconceptions over what errors they are able to monitor.

1.1 Why Do Pilots Need a Disciplined Check or Scan Regime?

Poor visual scanning signals a split between the pilot and the automated system that they are operating. The purpose of a systematic instrument check is to ensure that the pilot receives acceptably fresh information with a sufficient refresh rate to maintain a consistent level of SA. When a scan breaks down, a pilot's mental model of the aircraft's mode state or geospatial and temporal position is obsolete.

Errors or breakdowns in scanning and checks are invariably caused through: Insufficient capacity leading to reduction in Situational Awareness (SA); a distraction leading to subconscious re-prioritization; or, a fundamental misunderstanding of what is required. The latter, requires training and re-education and forms basis of this paper. The differing scan pattern failures can be split into four types. (1) *Incorrect scan*. This mis-interpretation of what is required may originate from poor initial training, a lack of competence, mental temporal or geo-spatial displacement from a loss of SA or that they have forgotten due to lack of recent exposure. It has also been shown that inexperience can lead to a change in scan pattern [3]. (2) *Degraded scan pattern*. All pilots are susceptible to this monitoring degradation and the brain is quick to prioritize what it considers worthy of attention in a scan [4]. Pilots who have not employed the scan in a while invariably scan at a slower rate than normal. (3) *Non-existent scan*. A complete failure to check a data source through distraction, re-prioritization or being unaware of the requirement. (4) *Insufficient or inappropriate scan*. When pilots drop elements of their scan, they often do not remember dropping them unless they are triggered to do so. Pre-conditioning from years of repetition, can also lead a pilot to think they have conducted a check when in fact they have not. They may think they recall the parameters, and in these cases they may only be recalling scanning the source; hence, pilots can gain a false impression of their own ability to monitor [4]. It has also been shown that the more inexperienced pilots are, the longer the dwell time they have with less regular fixations [1, 5] or less relevant fixations [6]. This can lead to the scan failing to be either efficient or effective largely because it leaves no time to perceive the data, let alone comprehend it [7].

There has been some thought as to why inexperienced pilots struggle in this manner. Airbus direct pilots to always use the highest level of automation available, and modern flight decks rarely require pilots to conduct raw data approaches. This can lead to a 'misuse' or over reliance of automation and, in turn, a poor scan technique [8]. Primary Flying Display (PFA) scan technique is not taught in many organizations whose aircraft may be a trainee's first exposure to a PFD. Unless they have trained on a modern glass cockpit aircraft, it is not possible to categorically know that the trainee has developed the correct scan technique. Moreover, with modern aircraft, the pilot arguably manages the system, rather than manually 'flies' it in the traditional sense. For this reason, it is essential that they know exactly what mode the airplane is in, something that can only categorically be known by a check of the mode annunciator.

To date the most researched function of pilot eye movement analysed is the scan pattern. Pilot scans can break down [8], and different pilots have differing scan rates [3, 9]. ET is a powerful tool that in these examples provided us with this information.

1.2 Identifying Poor Checking and Scanning in the Training Environment

When a student fails to carry out a manual action despite confirming that they have, it is clear to a vigilant instructor. When a student fails to check a data source despite having said that they have, it is less obvious. As they can only rely on head movement and perceived direction of gaze, a perennial and as yet unaddressed problem is that of instructors not being able to identify this failure. If there is no subsequent impact, this poor monitoring will go further unnoticed. This can detrimentally re-enforce bad practice through negative transfer [10], increasing the chance that the next time there is a mode or source of information contrary to that which is required, it will not be spotted. More often than not, as no adverse consequences transpire, pilots may not even realize their monitoring is inadequate [11]. Relaxation can develop, leading to normalized deviance from the correct SOPs.

Training a pilot involves the introduction of new information and techniques, followed by a period of exposure and practice. Part of the training process is to capture any poor techniques before they lead to errors. The check or scanning error that occurs due to a fundamental misunderstanding of what is required presents an insidious threat to flight. In some circumstances these issues cannot be identified with current training tools and the latent error may be repeated hundreds of times over many years, with little or no impact - until the right combination of triggers create an accident opportunity [12].

At present, using level D simulators, instructors are able to monitor and, through DBFs, graphically reproduce vast amounts of the simulation in real time. Unfortunately, DBFs are rare, but where they are available, the recorded session can be played back. The student will have a mental model of both their performance and what the correct performance should be; the difference between these two models may be further apart than the student realizes and verbal analysis may not bridge the gap. If the student understands the correct theory, video playback facilitates the student in contextualizing and adjusting their own actions in line with their own mental model.

1.3 Eye Tracking in Flying Training

Legacy, rudimentary methods of observing where a student is looking involve using a system of mirrors however, the analyst is not able to pinpoint exact fixation points, merely the rough direction of viewing [2]. It is only possible to identify an incorrect scan if parameters fall or remain outside limits. A system with the ability to monitor exact pupil movement has utility in the aviation training environment. We know that ET is a useful aid to help train pilots, but few studies address the manner in which this training should take place, providing practical, repeatable and employable techniques that can be incorporated into the simulator training suit. What aviation simulators and DBFs do not currently offer is the ability to monitor where a pilot is looking. Integrating ET data with the DBF for the purpose of identifying where the trainees are looking and enhancing the de-brief, would create an exceptionally powerful training tool.

1.4 Aim and Objective

This study integrates pilot ET data with an aviation simulator DBF. The aim was to investigate whether ET made a significant difference to the identification of pilot errors in ground-based pilot training. To do this the objective was to ascertain whether simulator instructors, both with and without the use of ET data, could identify whether a pilot was checking what they should be at the correct moment. Moreover, it was necessary to understand whether exposure to ET data changed instructors' pre-conceptions about how many errors they were able to see.

2 Method

2.1 Target Sample

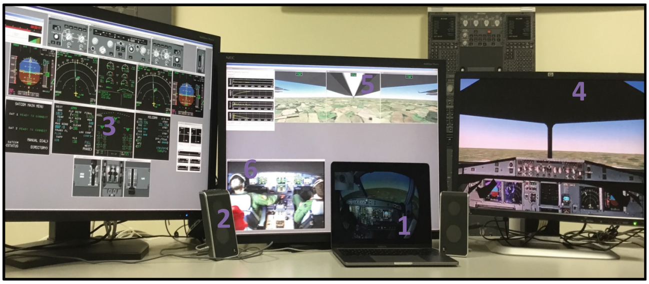
19 simulator instructors and examiners took part in the research. All subjects were either UK CAA Type Rating Instructors (TRI) or Examiners (TRE) or they were military Line Training and/or Air to Air Re-fueling instructors. The instructors were requested to support the research but were told no more about would be involved. From a total of 33 company instructors, the participation rate of 54.2%, easily ensured a probability of 0.95 that the number of subjects obtained was within ±0.1 SD of the true population mean. Demographic data: Gender (all men); age in years (M = 48.1, SD = 9.1); total flying hours (M = 9379.0, SD = 4708.7); total hours on an A330/Voyager (M = 2484.2, SD = 1790.8); and, total years as an A330/Voyager instructor (M = 6.38, SD = 6.01). Note: The Voyager is a military variant of the A330-200.

2.2 Experiment Hardware

Simulator. 3 pre-recorded profiles were made in a Thales, A330-200 Level D sim.

Eye Tracking. ET data was recorded using the Pupil Lab Eye Tracker.

Simulator DBF. The Thales DBF, shown at Fig. 1, comprises a desktop computer and combination of monitors and speakers which are linked to the flight simulator. The instructor is able to record parts or all of the simulator session, managed from their seat in the simulator. These recordings can be played back on the DBF. The DBF reproduces the flying displays, all sound and representations of the controls for the spoilers, flaps, thrust levers and landing gear. Additional information includes rearward facing camera images, flying control position data, SatCom data and a recording management window. A fourth monitor off screen to the right displayed an ATC radar picture of the aircraft. The research pre-recordings were stored on the DBF database.



1 – Eye Tracking Data
2 – Speakers
3 – Aircraft Instruments
4 – Flight Deck View
5 – Rearward Facing Cameras
6 – Simulator Camera View

Fig. 1. Thales De-Brief Facility, in Voyager configuration, in addition to eye tracking data.

Questionnaire and Measure. 26 questions required a mix of answers using Likert-type scale and free description. The questionnaire was answered in three Phases: 1 - pre-DBF exposure; 2 - post-DBF exposure (without ET augmentation); and 3 - post-DBF exposure (with ET augmentation). The 3 Phases of questions sought to establish the subjects' opinions on the efficacy of the simulator and DBF as environments in which to identify pilot errors and omissions.

2.3 Experimental Design

During the pre-recording of the 3 simulator profiles, an ET device, linked to a laptop controlling the ET program, was worn by the right-hand seat pilot (PM). At pre-briefed moments and contrary to Standard Operating Procedures (SOPs), the PM deliberately either avoided checking a necessary instrument or fixated for too long on one. For the data gather, each subject was seated at a table in front of the DBF screens. Subjects first answered Phase 1 of the questionnaire and watched three profiles on the DBF. Subjects then answered Phase 2 of the questionnaire and in turn watched the three profiles again, this time with ET data run in parallel from the same starting point. Subjects finally answer Phase 3 of the questionnaire. As the data available in each phase developed, certain question sets were repeated, capturing any changes of opinions.

ET was not mentioned until just prior to the ET exposure to avoid preconception bias and to ascertain the specific areas of instruction that each subject was focused on. The incorporation of the ET data was presented to mimic having it integrated into the DBF, see Fig. 1. The two exposures were the independent variables and the questions were the dependant variables. Figure 2 shows the additional data that ET can augment a DBF with.

3 Results

3.1 Challenging Pre-conceptions

The difference between '*How many of a pilot's actions the subjects thought they could monitor during simulation*', was tested using a repeated-measure ANOVA to examine instructors' responses in Phases 1, 2 and 3. It was found that there was no significant

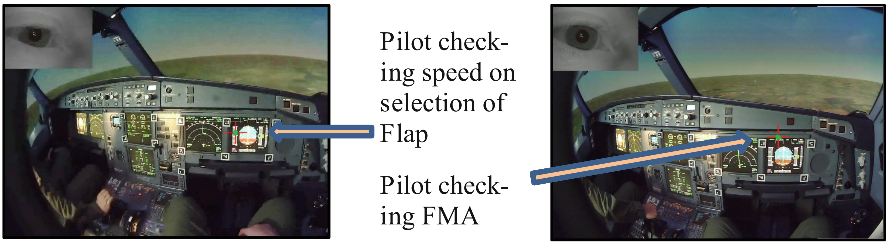


Fig. 2. Data available from ET.

difference between instructors' responses, F(2,36) = 0.368, p = 0.694, $\eta_p^2 = 0.02$. The null hypothesis that 'there is no significant difference among these three scenarios' can be accepted. Next, the difference between '*How many of a pilot's actions the subjects thought they could review in the DBF'*, was tested. A repeated-measure ANOVA was also conducted and it was found that there was a significant difference of instructor response to the 3 different Phases, F(2,36) = 8.348, $p < 0.05$, $\eta_p^2 = 0.317$. Phase 1–3 rating scale means and question decodes are at Fig. 3. The null hypothesis that 'there is no significant difference among these three scenarios' can be rejected. ANOVA results for questions relating to DBF review are at Table 1. Additionally, a post-hoc comparison by Tukey HSD indicated that there was a significant difference between Phase 1 (M = 6.95, SD = 1.43) and Phase 3 (M = 7.53, SD = 1.31). There was also a significant difference between Phase 2 (M = 6.47, SD = 1.47) and Phase 3 (M = 7.53, SD = 1.31), see Table 2.

Another repeated-measure ANOVA was conducted for responses to the three Phases to test the difference between questions 7, 12 and 12C. 12C was the answer that some instructors chose to change Q12 to when given the option in Phase 3. In other words, 'now they had been made aware of some of the errors and actions they had been missing, did they wish to revise their previous rating?'. This revision did not suggest ET augmentation was now available but sought to challenge instructors' pre-conceptions over their efficacy of their monitoring.

Again, there is a significant difference of instructor pilots' response, F(2,36) = 7.448, $p < 0.05$, $\eta_p^2 = 0.293$. The null hypothesis that 'there is no significant difference among these three scenarios' can be rejected. A post-hoc comparison by Tukey HSD indicated that there was significant difference between Phase 1 (M = 6.95, SD = 1.43) and Phase 3C (M = 5.79, SD = 1.62).

By showing subjects what they had missed, this challenged their pre-conceptions of their ability to identify error using the DBF, and changed their opinion.

3.2 ET as an Instructional Tool

When rating *how effective the subjects thought ET would be in improving an instructor's ability to monitor correct SOPs and scan patterns,* responses in Phase 2 (M = 7.37, SD = 1.21) rose in Phase 3 (M = 7.68, SD = 1.25) after exposure to ET data. Subjects then rated the efficacy of the ET data first as *a stand-alone tool*

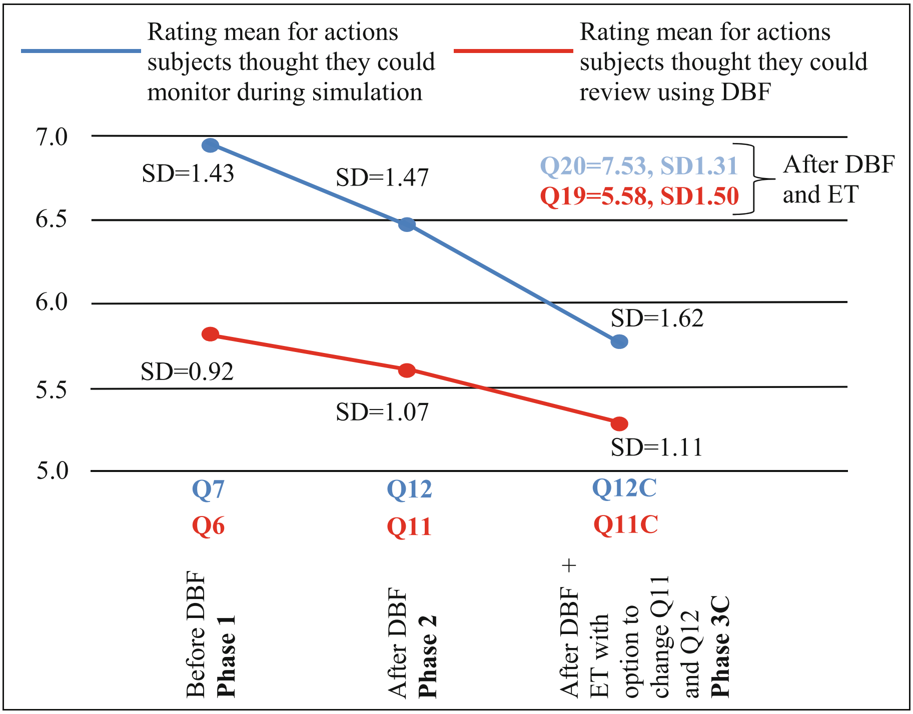


Fig. 3. Phase 1–3 rating scale Means for the number of a pilot's actions the subjects thought they could monitor during simulation and in DBF review.

Table 1. Summary of within-subjects ANOVA

Questions	SS	Df	MS	F	p	η_p^2
7, 12 and 20	10.561	2	5.281	8.348	0.001	0.317
Error	22.772	36	0.633			
7, 12 and 12C	12.877	3	6.439	7.448	0.002	0.293
Error	31.123	36	0.865			

Table 2. Summary of post-hoc Tukey HSD for comparable means.

Question pair	*t*-TEST				
Question (phase)	*t*	df	*p*	SE	Cohen's d
7(1) & **12**(2)	1.92	18	0.07	0.25	0.33
7(1) & **20**(3)	2.16	18	<0.05	0.27	0.42
12(2) & **20**(3)	4.06	18	<0.05	0.26	0.72
7(1) & **12C**(3)	4.01	18	<0.05	0.29	0.76
12(2) & **12C**(3)	1.91	18	0.07	0.36	0.44
15(2) & **25**(3)	0.40	18	0.69	0.39	0.25

(M = 6.68, SD = 1.60) and then *when used in conjunction with the DBF* (M = 7.89, SD = 1.29), a notable increase. Finally, subjects were asked a standalone question to assess *how effectively they thought they were at integrating ET data with the DBF data*, (M = 6.79, SD = 1.03).

Pilot Error Analysis. Errors identified by the subjects in profiles 1, 2 and 3 were noted down in Phases 2 and 3. Across all subjects, 42 separate error types were identified from a total of 93 individual observations, before the addition of ET data. Errors included 'not setting Missed approach alt', 'not checking approach lane prior to line up' or 'Side Stick (SS) in wrong position at take-off'. After the introduction of ET, 23 additional error types were identified from a total of 88 individual observations. Notably, from the 23 additional errors identified in Phase 3, only one was also captured in Phase 2. Out of a total of 64 separate errors, 22 new errors (34.4%) were identified with ET augmentation. The number of error observations in Phase 2 (M = 4.74, SD = 1.58) was 11.3% of the Phase 2 total observations rising in Phase 3 (M = 5.84, SD = 2.33) to 25.4% of the Phase 3 total. It could also be seen that every new error identified could be directly attributed to the availability of additional ET data.

Pilot Action Analysis. Pilot actions that instructors reported not being able to see were captured both before and after the introduction of ET from 52 individual observations. A total of 15 separate actions were identified across all subjects before the addition of ET data and 9 separate actions after the introduction of ET. Only 2 of the 9 actions from Phase 3 were mentioned in Phase 2. From this we see that out of a total of 24 separate actions, 7 new actions (29.1%) were introduced following the ET augmentation. Of the 15 separate actions 'not seen by the instructor' in Phase 2, 13 of the 31 observations (42%) would have been remedied by ET. Additionally, two subjects did not consider that there were any actions that they did not see, although when they had seen the ET they were able to identify some. In Phase 3, having seen the ET, there were 21 additional observations added, of which 7 actions 'not seen by the instructor' had not previously been mentioned. This showed that following exposure to ET, instructors further identified an additional 47% actions not seen. Of these, 71% would be solved by ET.

Improving the Current Information Sources. Table 3 shows the methods that instructors believed would increase the level of relevant information available to them, as noted before the introduction of ET data. The actions highlighted orange are those that ET would remedy, 45.5% of the total methods.

4 Discussion

ET has been successfully implemented in training before [9], however this study was able to robustly support the hypothesis that ET made a significant difference to an instructor's identification of pilot errors in ground based flying training. This paper does not attempt to address more specific outputs from ET data such as dwell time, lack of fixations or random scan patterns. There has been much research on these subjects and it is known that pilot weakness in these areas can be identified through ET. What

Table 3. Methods to improve information availability.

	Methods	Observations
1	Reproduce SS inputs on instructor screen	3
2	Cameras at side of cockpit to monitor eye activity	2
3	Cameras at side of cockpit to monitor SS activity	2
4	Reproduce FMGC inputs on instructor screen	2
5	Camera looking at pilot	2
6	Eye tracking	2
7	Ability to view body language	1
8	Reproduce rudder inputs on instructor screen	1
9	Knowing what the pilots are checking when responding	1
10	More camera angles	1
11	Use of DBF to show NOTECH issues	1

appears to have been ignored is understanding whether instructors are able to process, interpret and utilize this data; moreover, whether its integration within DBFs is accessible and of positive benefit.

Challenging Pre-conceptions. Q6, 11 and 19 related to errors identified during simulation. Whilst showing the subjects DBF data directly affected their opinion relating to Q7, 12 and 20, this opinion had to be inferred when considering simulation – this is discussed further under ‘ET as an instructional tool’. For Q6, 11 and 19, ratings indicated that, even with the exposure to the DBF, subjects were content with their assessment of their ability to spot errors. It was assumed that with the additional ET exposure in Phase 3 and having noted down the considerable numbers of additional errors that they missed in Phase 2, they would again reduce their rating assessment of the errors they were able to identify. Instead, the Phase 3 mean stayed constant. This demonstrates that the ET data exposure in the DBF did not alter the instructors’ pre-conceptions of what errors they were able to identify in the simulator.

Q7, 12 and 20 related to errors identified using the DBF. Having conducted the research it was ascertained that many of the subjects did not routinely use the DBF for their instructional de-brief. For some, they were contractors that also instructed for other companies. A DBF is a rare commodity and they only usually got access to one at RAF Brize Norton. Others did not consider that the DBF added enough weight to a de-brief to justify its setup and use. For these reasons, they were not practiced in its functionality and refrained from using it. In conversation only around 5 (26%) of the instructors stated that they used it regularly, notably those with more experience. As it was known that so few instructors used the DBF to augment de-brief, it was necessary to track their evaluation of its worth, both before and after seeing it in action. The mean rating reduction of 0.48 between Phase 1 and 2, t(19) = 1.92, p = 0.07, indicated that instructors felt they were now less able to identify errors, having just seen the DBF in action. An explanation for this fall is that on exposure to the DBF and the excess of information that is available, instructors felt unable to monitor all the data at once and were thus prone to missing errors. By adding ET data, subjects' mean rating score increased from Phase 2 by 11.8%, to 84.7% of maximum rating, evidence that instructors recognized and accepted missed errors that they had previously been unaware of. This strongly validated the positive effect that ET has in helping to identify errors in training. When given the option to revise their Phase 2 rating, based on what they now knew, Phase 1 to Phase 2 gave a statistically significant drop; t(19) = 4.01, p = <0.05. This change demonstrated that the exposure to ET data significantly altered the instructors' preconceptions of what errors they were able to identify using the DBF.

Eye Tracking as an Instructional Tool. Instructors rated the use of ET as a standalone instructional de-brief tool, when considering it to be the sole source of information. This question assessed the value the subjects gave to the ET data. A 74% mean rating response rose to an 88% when consideration was given to ET being used in conjunction with the DBF i.e. Phase 3. It can be seen that instructors assessed that the combination of both sources of data was a powerful training tool.

Instructor Error Awareness. It has been shown that the instructors' awareness of the limitations of the simulator and DBF are broadly aligned. In Phase 2 the number one perceived issue related to SS input visibility. This problem, unique predominantly to Airbus, was a contributory factor during the Air France 447 A330 crash, where neither pilot could see each other's SS, exacerbating a lack of awareness of their opposing inputs [13]. Although not related to ET, it is interesting to note that detractors of Airbus often cite the SS as being a negative of the cockpit design and the most high-profile Airbus accident in the last few years was linked to an inability to see SS inputs. It is quite possible that that this knowledge drives the SS issue to the forefront of instructors' minds. The issue with the most overall observations related to tracking of gaze and scanning. Even before the introduction of ET data, the number of 'cannot see where eyes are looking' observations, in addition to the many other associated eye gaze issues, demonstrated that not being able to see what the trainee is looking at is a concern within the instructor cadre.

Error Response. The unusual scenario of watching the DBF without first having sat through the simulator profile presented some compelling data. Phases 2 and 3 exposed all instructors to exactly the same profiles. In phase 2 the instructors were not able to see the scanning and checking errors, therefore these was not necessarily their focus. Their standard brief asked them to identify the recorded pilots' compliance with SOPs, leading them to seek other mistakes. During the pre-recording and despite their best efforts, minor errors were made by the pilots in both setting up the flight deck and performing the profiles. In Phase 2, each subject identified a mean of 4.74 (SD = 1.58) out of a total of 42 errors, demonstrating that instructors seek different sources of information when un-prompted. It shows that different things are important to different people at different times. It is difficult to explain these varying foci however, they would suggest that the multiple sources of information that individuals have exposure to in their daily life have mentally primed them differently. They potentially have a cognitive bias towards certain errors and hence seek them out when their attention is not targeted elsewhere. Support for this came when interrogating actions that instructors did not think they could see; their primary focus in Phase 2 was on the SS, potentially due to the Air France 447 accident twinned with Airbus' use of a SS. From the number of observations, we also see that in Phase 2, where there is no specific focus to the error identification, the ratio of observations to separate error types is 2.21 to 1. After the introduction of ET, where error identification becomes focused, the ratio of observations to separate error types is 3.83 to 1. This clearly demonstrates that with targeted training, we can align instructor focus and increase identification of error.

5 Conclusion

This research conclusively shows that using ET allowed instructors to spot increased numbers of errors. As trainees look forwards in the simulator, their face generally cannot be seen. When they incorrectly scan and poorly monitor their errors cannot be spotted by an instructor if no subsequent failure or omission occurs. Integrating ET information was considered challenging due to the additional volume of data however, it is thought to significantly improve the DBF capability, creating additional training opportunities for students. It also shown that if conducting training using the DBF, the target of instruction must be focused otherwise error identification is random. ET's key advantage however, is that is measurably focuses and enhances instructor attention on identifying checking, scanning, and monitoring errors. Observing the DBF, both with and without ET augmentation, increases instructor understanding of what they are unable to identify in both simulator and DBF. ET even changed their pre-conceptions regarding the efficacy of their trainee monitoring, reducing their levels of false confidence and educating them on how they could be better.

References

1. Bellenkes, A.H., Wickens, C.D., Kramer, A.F.: Visual scanning and pilot expertise: the role of attentional flexibility and mental model development. Aviat. Space Environ. Med. **68**, 569–579 (1997)
2. Ferrari, F., Spillmann, K.P., Knecht, C.P., Bektas, K., Muehlethaler, C.M.: Improved pilot training using head and eye tracking system. In: ET4S 2017 (2018). https://doi.org/10.3929/ethz-b-000225616
3. Sullivan, J., et al.: Training simulation for helicopter navigation by characterizing visual scan patterns. Aviat. Space Environ. Med. **82**(9), 871–878 (2011). https://doi.org/10.3357/ASEM.2947.2011
4. CAA: Flight-crew human factors handbook CAP 737, p. 242 (2014)
5. Matessa, M., Remington, R., Field, M.: Eye movements in human performance modeling of space shuttle operations, pp. 1114–1118 (2005)
6. Kasarskis, P., et al.: Comparison of expert and novice scan behaviors during VFR flight. In: Proceedings of the 11th International Symposium on Aviation Psychology, January 2001, pp. 1–6 (2001)
7. Endsley, M.R.: The challenge of the information highway. In: Proceedings of the Second International Workshop on Symbiosis of Humans, Artifacts and Environment, September 2001

8. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. Hum. Factors: J. Hum. Factors Ergon. Soc. **39**(2), 230–253 (1997). https://doi.org/10.1518/001872097778543886
9. Wetzel, P.A., Anderson, G.M., Barelka, B.A.: Instructor use of eye position based feedback for pilot training. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 42, no. 20, pp. 1388–1392. SAGE Publications Inc. (1998). https://doi.org/10.1177/154193129804202005
10. Harris, D.: Human Performance on the Flight Deck. Ashgate Publishing Limited, Farnham (2011)
11. Dismukes, R., Berman, B., Loukopoulis, L.: The Limits of Expertise. Ashgate Publishing Limited, Farnham (2007)
12. Reason, J.: Human Error. Cambridge University Press, New York (1990)
13. Bureau d’Enquêtes et d’Analyses pour la sécurité de l’aviation civile: Air France 447 Final Accident Report (2012). http://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf

Conceptual Modeling of Risk Interactions for Flight Crew Errors in Unmanned Aerial System Operation

Yi Lu[1(✉)], Huayan Huangfu[1], Shuguang Zhang[2], and Shan Fu[1]

[1] Department of Automation, School of Electronic Information
and Electrical Engineering, Shanghai Jiao Tong University,
800 Dongchuan Road, Shanghai 200240, China
luyil@sjtu.edu.cn
[2] School of Transportation Science and Engineering, Beihang University,
37 Xueyuan Road, Beijing 100191, China

Abstract. As revealed by the mishap causal factor statistics, flight crew error was the most critical factor to affect the safe operation of large Unmanned Aerial System (UAS). However, most research on this topic embedded their roots on event-chain based accident model or even relied on textual descriptions to discuss the risk mechanism leading to large UAS mishaps. It is hard for preventing future losses of UASs in a systemic and efficient way, especially consider the organizational context of flight crew error shaping factors. Based on the System Dynamics approach, this study proposed to illustrate the risk mechanisms that involve the interactions of organizational, flight crew and technical system factors leading to operation accident of large UASs. It aims to explain such risk interactions and helps to clarify why flight crew training always gained a delayed safety benefits on crew error reduction, especially when UAS fleet facing high mission tempos. This study also discusses the effect of flight crew self re-learning process and safety commitments on UAS operation accidents. Some Causal Loop Diagram (CLD) based conceptual models are established for future quantitative SD stock-flow simulation which can be used as organizational risk assessment tools, when considering to evaluate the medium-long term benefits of potential safety policy and management decision in the widely spreading UAS operations.

Keywords: Large UAS operation · Risk interactions · Flight crew errors · System dynamics

1 Introduction

Unmanned Aerial System (UAS) is constructed by the Unmanned Aerial Vehicle (UAV), ground control station, data links and recovery & launch system [1]. Derived from their military cousins, the usages of large UAS spread rapidly since 1990s (referring to the Category III specified by US DOD and FAA, the large UAS is the category with the maximum take-off weight above 599 kg [2, 3]. Meanwhile, due to the inherent advantages on low-cost compared to manned aircrafts, the civil large UASs

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 271–283, 2019.
https://doi.org/10.1007/978-3-030-22507-0_22

have vast potentials for future development, such as coast guarding, geology exploring, filed investigate, etc. According to statistics since 2010, the size of global civil UAS market has grown to $100 billion and was growing year by year and the UAS regulator and industry were considering the integration of large UAS into national airspace [4, 5]. However, according to the statistics of the US Office of the Secretary of Defense (OSD) in 2003, the average Class A mishap rate per 100,000 h of the military UASs in worldwide was an order of magnitude higher than the manned aircraft [6]. From 2004 to 2006, 20% of Class A mishaps (i.e., causing the total loss valued above $100 million or causalities) of the US Air Force (USAF) can be attributed to the MQ-1 Predator fleet which had 21 mishaps in total and 17 vehicles completely destroyed. Moreover, in the single year of 2015, the MQ-1 Class A mishaps comprises about 57% percent of the total Class A mishaps of the USAF and a growth acceleration of the scale can be seen. Most importantly, the current safety level of civil UASs cannot satisfy the airworthiness requirement of 1 Class A mishap per 100,000 h, which challenges the UAS safety engineering in the future and attracted more attentions around the global.

In order to ensure the sustainability of aviation growth, researchers have been conducting statistical analyses over aviation accident risk factors since 1990s, realizing an overall safety trend: with the increasing of operation frequency and the accumulation of operation time, the reliability of technical system was continuously improved, which induced the accidents caused by human and organizational factors to become more and more obvious. More importantly, this trend was more pronounced due to: (1) the limited field of pilot view and spatial cognition have induced a large number of flight crew errors, the ratio of this causal factor is 50% higher than in manned aircraft; (2) the high-strength tasks have intensified the flight crew error to be one of the most common causes of UAS accidents; (3) the handing transfer between flight crews (i.e., LRE/MRE) introduced risks on operation when considering the emergency treatment [7, 8]. Facing this, theories on risk mechanisms involving non-technical factors were raised and have been applied on the UAS accident analysis and safety improvement by using the event-chain model as a framework, such as the Human Factors Analysis and Classification System (HFACS) [9, 10]. HFACS has gained the benefits on decreasing the aviation accident rate. Using the USAF MQ-1 Predator fleet as a case, after the application of the HFACS in 2001, its Class A accident rate of cumulative 10,000 flight hours changed from 43.9 (2001) to 8 (2011), however, it maintained a stable level of 7 to 8 thereafter. Such trend shows: although the probability-based risk theories truly have the initial effects on risk reduction, their component-failure based view-point still embedded roots in static and linear safety philosophy and can hardly address accident causality involving interactive complexity (technical, organizational and human). Meanwhile, some other researchers have made efforts to develop extension techniques to investigate human reliability factors, such as Petri nets, Dynamic Bayesian Network and statecharts [11]. However, these methods can not address the dynamic processes of risk transferring, especially in human and organizational levels. With view of systems theories, the development of working environment, process, and infrastructure that enables the human factors considerations to support the success of operation in the long-term and human factors sustainability is important especially in the field of aviation.

Properly understanding the risk interactions in UAS operation process requires first understanding how and why this social-technical system migrate towards states of increasing risk. By identifying these risk mechanisms in a causation model based way, UAS operation organizations can better understand past accidents, monitor risk, and decrease the likelihood of future losses by identifying UAS operation risk spectrum in different levels. Moreover, in accident prevention, compared to those textual descriptions as products of traditional UAS mishap investigation, such risk spectrum can be developed as quantitative tools to evaluate the medium-and-long term benefits of organizational safety investments.

2 Brief Overview of System Dynamics

Grounded on the theory of nonlinear dynamics and feedback control, System Dynamics (SD) is a framework for dealing with dynamic behavior of complex system and also draws on cognitive and social psychology, organization theory, economics, and other social sciences [12]. In the field of system safety, system dynamics has been used as an important supplement to analyze organizational accidents and proposed safety policy in the field of aviation, astronautics and chemical industries [13–15]. Especially, in a view of social-technical system, some researchers began to use SD for organizational accident analysis [16, 17]. SD helps to model the risk interactions of organization safety with conceptual description, causation analysis and time-domain simulation tools. The risk interactions can be described by the use of three basic SD modeling elements: the reinforcing loop, the balancing loop, and the delay.

2.1 Reinforcing Loop

It refers to a particular behavior that encourages similar behavior in the future and it corresponds to a positive feedback loop in control theory. As Fig. 1 shows, an increase of Variable A causes a positive consequence in Variable B $(A \xrightarrow{+} B)$, as indicated by "+", which then also causes an increase in Variable A (i.e., R-loop). For an example of positive consequences, the increase of training investment can improve the flight skills of Launch and Recovery Element crew. The R-loop can also be applied to negative consequences (i.e., indicated by "−").

2.2 Balancing Loop

It exists when a particular behavior attempt to move from a current state to seek balance. It corresponds to a negative feedback loop in control theory, as the B-loop (A-C-B-A) in Fig. 1 shows. The driving force in the loop is the size of gap (i.e., Variable B) between the goal (i.e., Variable A) and current value (i.e., Variable C). For example, facing the gap between actual technical system reliability (it is limited by national industrial level) and its goal (it is required in design specification), both the design modification and emergency procedure revision (to deal with unexpected incident involving system failure or malfunction) will be promoted, which help to reduce such gap and gain a relevant balanced UAS safety level.

2.3 Delay

It is used to model the time the actions need to take effect and may result in unstable system behavior. It is indicated by a double line as shown by the relationship between Variable C and B in Fig. 1 (i.e., $C \xrightarrow{\text{Delay (-)}} B$). Caused by the delays, actions are deemed unsuccessful prematurely to achieve expected results. For example, due to flight crew mission experience needs time to accumulate, operation organizations always obtain a delayed training benefit compared to their investments. It should be noticed that delays can occur within both balancing and reinforcing loops.

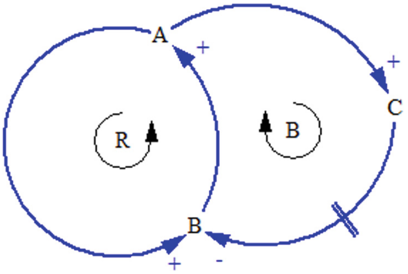


Fig. 1. Reinforcing loop, balancing loop and delay.

2.4 Causality Loop Diagram

In this study, the critical risk interactions involving UAS development-operating-maintenance (DOM) processes are modeled with the language of Causality Loop Diagram (CLD) which is used as conceptual modeling tools in SD [12]. The reinforcing and balancing feedback loop (R/B) structure provides a framework for loop dominance analysis which explains UAS operation risk mechanism. The data supporting the establishment of CLD include:

(1) Engineering assumptions grounded in practical experience and accident investigation related to DOM processes flaws.
(2) Organization behavior modes and safety features proposed in literatures reviews, such as accident and risk models.
(3) Accessible UAS operation data, such as flight crew error specified in accident statistics, training investment records etc.

In this study, VENSIM software developed by Ventana Systems is adopted to illustrate the proposed CLDs due to this software provides convenient PLE version for educational and academic use. The detailed features of this software can be found in its reference manual.

3 Data Sources and Analysis Methods

3.1 Data Sources

The UAS have a broad spectrum of type and operation field. The early operation of UAS was implemented by military planners to carry out reconnaissance and/or attack missions. Thanks to UAS's convenience and low-cost characteristics, the use of UAS was rapidly expanding to more civil domains. They now have outnumbered military UAVs vastly, with estimation of over several millions per year. However, the widespread UASs threaten airspace security in numerous ways, including unintentional collisions with population and other manned aircrafts. Beware of UAS's convenience and low-cost characteristics, the use of UAS was rapidly expanding to more civil areas, such as disaster relief, environmental conservation, filmmaking or cargo transports, and they now have outnumbered military UAVs vastly, with estimates of over several millions per year. Indeed, in terms of quantity, most members in civil drone family are quadcopter types or short scale fixed wing designs operated under direct visual line of sight, which have relatively low impact energy and limited remote range. In contrast, remained as an overarching concern for most Aviation Authorities worldwide, the most significant risks to public safety come from the operations of those large UASs featured by bigger size/weight, long-endurance, high speed and payload and the majority are military types even derived civil types with similar configurations, such as the US Air Force MQ-1 Predator series [18].

Based on those open sources, many causation analyses of large UAS accident were implemented. For example, Tvarynas et al. found the frequency of human factors mishaps of US military large UAS fleet was increasing, with data on UAS mishaps during fiscal years 1994–2003 [18]. Nullmeyer et al. used the USAF MQ-1 Predator Class A mishaps as a case study and derived flight crew training measures [7, 8]. Consequently, MQ-1 Class A mishaps attribute to human error (flight crew and maintainer) decreased despite increasing numbers of mishaps overall. Especially, USAF Sustainment Center generates mishap investigation reports for typical UAS for every Class A mishap by fiscal year and by UAS type and provides results at varying levels of granularity. Based on such detailed statistics, the safety records of the USAF MQ-1 fleet were summarized in this study and the Class A mishap contributors from FY 1996 to 2017 were specified as shown in Table 1 and 2. As an accessible case, it provides the basic operation data for establishing a general model to describe the UAS operation safety, emphasizing the flight crew error related risk interactions.

Table 1. USAF MQ-1 Class A mishap statistics by flight phase (FY 1996–2017).

Flight phase	Take-off	Climb	Landing	Cruise	Approach	Go-around
Proportion %	10.5	9.4	12.7	47.5	15.3	4.6

Table 2. USAF MQ-1 Class A mishap statistics by causal factor (FY 1996–2017).

Causal factors	Propulsion system failure	Flight control system failure	Data link failure	Flight crew error	Maintenance personnel error	Other items
Proportion %	24.3	8.5	7.9	27.8	25.4	6.1

3.2 Framework of Risk Analysis

As indicated by literatures and high-risk organization accident investigation reports, human factors in aviation are a complicated concept including human physiology, psychology (perception, cognition, memory, social interaction, error, etc.), work place design, environmental conditions, human-machine interface and anthropometrics. It can also be divided into individual factors and group cooperation factors (i.e., Crew Resource Management, CRM). Based on the Swiss Cheese Model (SCM), the Human Factors Analysis and Classification System (HFACS) has built a bridge between theory and practice by accommodating human factors in aviation in a more systematic way [9]. The model considers all aspects of human errors, including the conditions of operators and organizational failures. It divides aviation accident related human factor into four categories which form a hierarchical structure, namely Unsafe Acts, Preconditions for Unsafe Acts, Unsafe Supervisions, and Organizational Influences. In this study, to identified UAS maintenance safety related risk factors, the HFACS level is adopted to determine the source of factors and the logic sequences beneath them, as Fig. 2 shows. Moreover, we also consider the risks of technical system level in this framework.

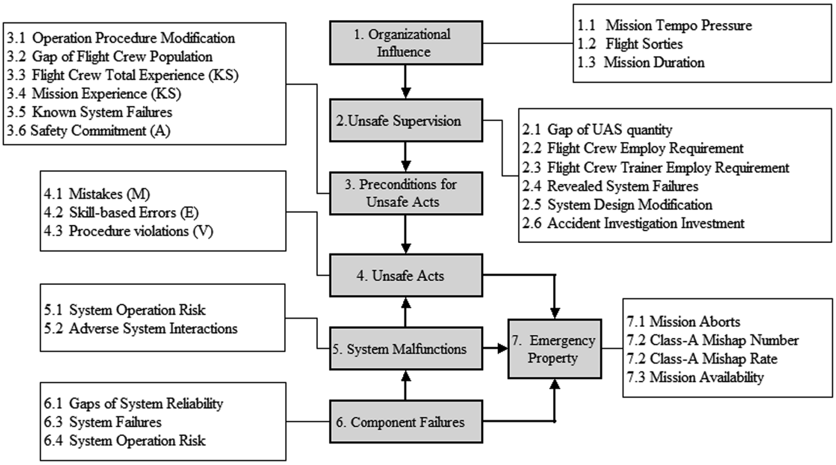

3.1 Operation Procedure Modification
3.2 Gap of Flight Crew Population
3.3 Flight Crew Total Experience (KS)
3.4 Mission Experience (KS)
3.5 Known System Failures
3.6 Safety Commitment (A)
1. Organizational Influence
1.1 Mission Tempo Pressure
1.2 Flight Sorties
1.3 Mission Duration
2.Unsafe Supervision
2.1 Gap of UAS quantity
2.2 Flight Crew Employ Requirement
2.3 Flight Crew Trainer Employ Requirement
2.4 Revealed System Failures
2.5 System Design Modification
2.6 Accident Investigation Investment
3. Preconditions for Unsafe Acts
4.1 Mistakes (M)
4.2 Skill-based Errors (E)
4.3 Procedure violations (V)
4. Unsafe Acts
5.1 System Operation Risk
5.2 Adverse System Interactions
5. System Malfunctions
7. Emergency Property
7.1 Mission Aborts
7.2 Class-A Mishap Number
7.2 Class-A Mishap Rate
7.3 Mission Availability
6.1 Gaps of System Reliability
6.3 System Failures
6.4 System Operation Risk
6. Component Failures

Fig. 2. Factors identification framework for UAS operation risk analysis.

In this table, the category item 3.3 and 3.4 represents the knowledge, skill and attitudes of flight crew which determine the proficiency of maintainer when implementing required tasks, e.g., replace system components following specified intervals and technical procedures. Here the flight crew involve the Launch and Recovery Element (LRE) and Mission Control Element (MCE) crews. Moreover, the category item *4.1 Mistakes* include two types, procedure and knowledge based mistakes. And the category item *4.2 Skill-based Errors* (derived from personal inability and poor training benefits) and *4.3 Procedure violations* also have their respective components.

This trend shows: although the probability-based risk theories truly have the initial effects on risk elimination, their component-failure based view-point still embedded roots in static and linear safety philosophy and can hardly address accident causality involving interactive complexity. Of the more than 100 Class-A mishaps occurring during the period of fiscal years 1996–2017, 47.5% happened in cruise phase and 28 mishaps involved flight crew errors. In fact, with the traditional statistics method of accident causal factors, it is hard to distinguish the different failure causes rooting in the flawed system design and/or inadequate operation activities. For example, An USAF MQ-1B Predator met electrical malfunction on Aug. 22, 2012, which led to the crash of the aircraft in a non-residential area in Afghanistan. The mishap investigation found the mishap UAS experienced an electrical malfunction due to a dual alternator failure, which began a chain of events that caused the aircraft to function solely on battery power [19]. In this process, the MCE crew failed to apply a checklist procedure that would have preserved more battery power. The MCE crew initialized an emergency mission abort procedure and handled over the control to the LRE crew when the recovery window opened, but the batteries were exhausted to the point where it was impossible for the aircraft to reach the runway. For this mishaps, causal factors were classified as belonging to both "propulsion system failure" and "flight crew error". In this term the causality revealed by the accident investigation was ignored and discounted information that can provide valuable experience for further system design modification and crew training improvement. In order to analyze the dynamic risk interactions related to flight crew errors and especially the behavior shaping contexts, the system dynamics approach is introduced to analyze the UAS operation risk mechanisms in terms of conceptual feedback loops.

4 Conceptual Modeling of Risk Interactions

Based on the data source and proposed framework of UAS operation risk analysis in Sect. 3, this section models the flight crew error related risk interactions with the system dynamics approach. In order to present the derived CLDs clearly, the whole model was divided as two views: (1) view of the emergency and organizational level, which mainly involve the risk factors of Categories 1, 2 and 7 shown in Fig. 2; (2) view of the crew resource and technical system level, which mainly involve the risk factors of Categories 3 to 6 shown in Fig. 2.

4.1 Emergency and Organizational Level

On the top level of large UAS operation risk dynamics model, the Emergency Level (EL) represents the indicators of UAS safety and mission availability (e.g., mission aborts caused by flight crew issues). In this level, some risk factors form the decision base or activity goal for organization management so the factors in the Organizational Level archetype (OL) are involved in one frame for better representation of their interactions, as shown in Fig. 3. In this figure, four balancing loops (B_1–B_4) and one reinforcing loop (R_1) are identified. They can be categorized into three groups as following:

(1) R_1 (TL_2-OL_5-EL_2-TL_2) and B_1 (OL_5-EL_1/EL_2-TL_2-OL_5)
In this group, the critical node variables are TL_2 and OL_5. The reinforcing loop R_1 take over the loop dominance at the early phase of risk interactions. In this means, the system failure induced mishaps in take-off and landing phases always occur at the initial period of a certain UAS type's service experience (i.e., mishap vehicle under control of LRE crew can be observed with a dimension of mission sortie). If considering the interaction between flight crew and system failure, such situation will be much worse. Meanwhile, the balancing loop B_1 take over the loop dominance later on. It means the contribution of mishaps in cruise phase appears with a delay.

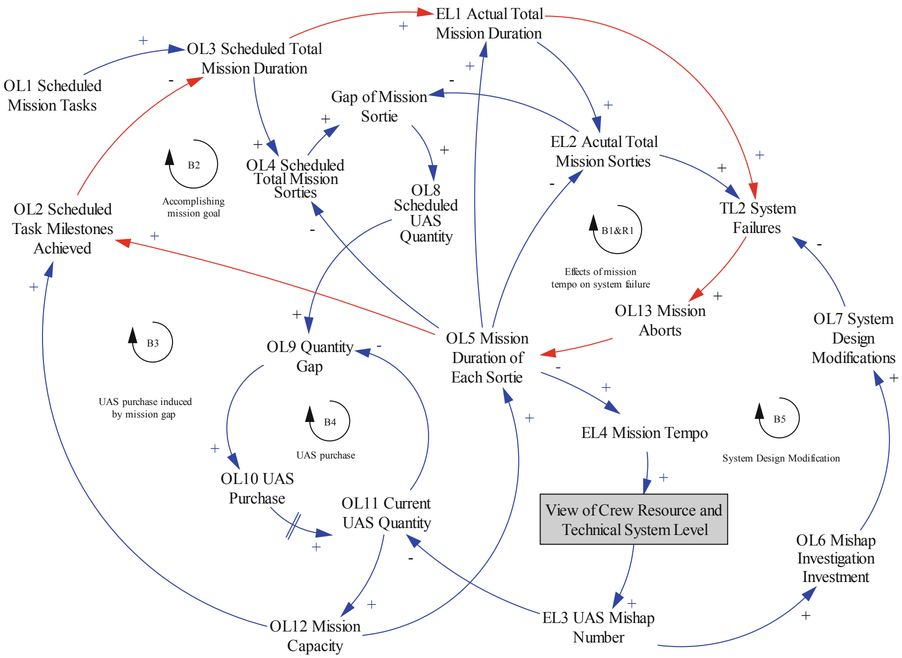

EL1 Actual Total Mission Duration
OL3 Scheduled Total Mission Duration
OL1 Scheduled Mission Tasks
Gap of Mission Sortie
EL2 Acutal Total Mission Sorties
OL4 Scheduled Total Mission Sorties
OL8 Scheduled UAS Quantity
B2
Accomplishing mission goal
OL2 Scheduled Task Milestones Achieved
TL2 System Failures
B1&R1
Effects of mission tempo on system failure
OL13 Mission Aborts
OL7 System Design Modifications
B3
OL9 Quantity Gap
OL5 Mission Duration of Each Sortie
UAS purchase induced by mission gap
B4
UAS purchase
EL4 Mission Tempo
B5
System Design Modification
OL10 UAS Purchase
OL11 Current UAS Quantity
View of Crew Resource and Technical System Level
OL6 Mishap Investigation Investment
OL12 Mission Capacity
EL3 UAS Mishap Number

Fig. 3. UAS operation risk CLD model in the emergency and organizational levels

(2) B_2 (OL_2-EL_1-TL_2-OL_5-OL_2)
Facing the increasing of system failures (EL_1) in operation, the system reliability problems (TL_2) influence the mission tempo (OL_5) and the organization always choose to take risks to achieve scheduled mission duration. This explain the underlying cause that how UAS mishaps may influence the availability of large UAS fleet in a long-term vision.

(3) B_3 (OL_2-OL_4-OL_9-OL_{12}-OL_2) and B_4 (OL_9-OL_{11}-OL_9)
In this group, the critical node variables are OL_9 and OL_{12}. They describe the organizational activity that purchases new UAS to compensate the mission capacity affected by UAS mishaps. Meanwhile, the organization also increase the scheduled mission sorties (OL_4) to compensate the mission capacity under reduced UAS quantity. The reinforcing loop B_3 explains such vicious spiral phenomenon which often appeared in the USAF MQ-1 Predator UAS operation history and were revealed by some accident investigation reports.

(4) B_5 (TL_2-OL_5-EL_3-TL_2)
This balancing loop describes the source of hindsight safety efforts raised by organizational accident investigation which plays an important role in connecting the communication between UAS development and operation processes through the system design modification, such as introduce fail-safe characteristics and improve component quality in technical system level.

Under high mission tempo, the risk interactions between crew resource and technical level (this potential view is indicated by a grey block in Fig. 3) play an important role in determine the UAS mishap number (EL_3), which also establish the internal relationships among the whole social-technical system carrying out the UAS operation process. More importantly, as the core factor in the view of this level, the variable "Mission Duration of Each Sortie (EL_5)" is adopted to reflect the actual task load of a single UAS, which defined the fuzzy term "mission tempo" in a quantitative way and made the potential SD simulation possible instead of a general textual description.

4.2 Crew Resource and Technical System Level

In this level, the risk interactions between crew resource (HL) and technical system level (TL) are focused. This CLD modelled their causations in two aspects: (1) the experience formation process effects of flight crew; (2) the mechanism of risk factors inducing flight crew errors. As shown in Fig. 4, two balancing loops (B_1–B_2) and four reinforcing loops (R_1–R_4) are identified. They can be categorized into two groups as following:

(1) B_1 (HL_6-TL_4-HL_6) and B_2 (HL_1-HL_8-EL_4-HL_4-HL_1)
This group induced the balancing loop B_1 to model the variation process of flight crew population in a simplified way. As the goal of this loop, the flight crew employment requirement (HL_7) is determined from the mission tempo (EL_4, its definition has been described above). Meanwhile, the population gap enforces the organization to employ more flight crew trainer (HL_3) to ensure the crew proficiency. Whether the flight crews can get adequate training resource determines their average experience which is also an

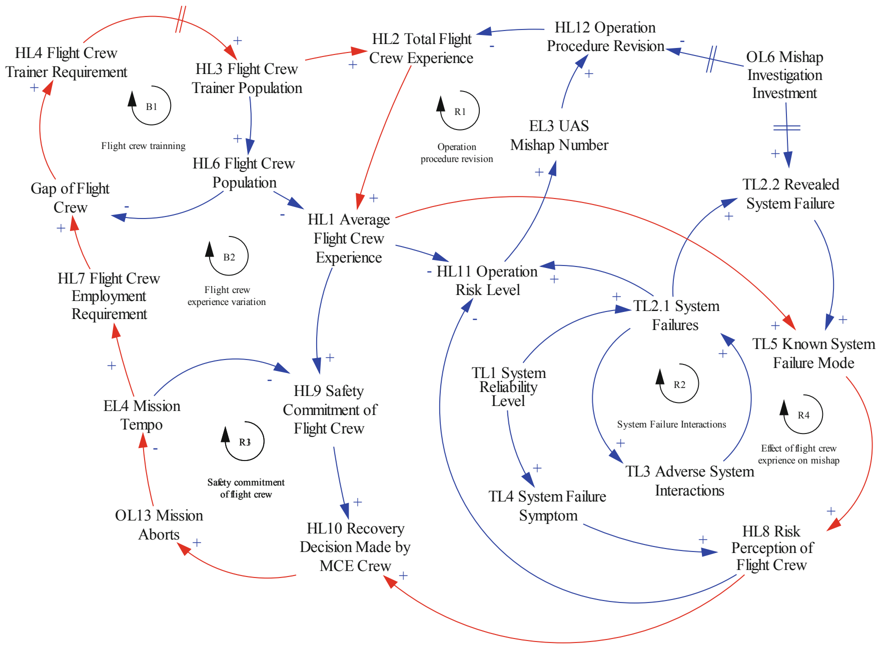

HL4 Flight Crew Trainer Requirement
HL3 Flight Crew Trainer Population
HL2 Total Flight Crew Experience
HL12 Operation Procedure Revision
OL6 Mishap Investigation Investment
B1
Flight crew trainning
R1
Operation procedure revision
EL3 UAS Mishap Number
HL6 Flight Crew Population
Gap of Flight Crew
TL2.2 Revealed System Failure
HL1 Average Flight Crew Experience
B2
Flight crew experience variation
HL7 Flight Crew Employment Requirement
HL11 Operation Risk Level
TL2.1 System Failures
TL5 Known System Failure Mode
TL1 System Reliability Level
R2
System Failure Interactions
R4
Effect of flight crew exprience on mishap
EL4 Mission Tempo
HL9 Safety Commitment of Flight Crew
R3
Safety commitment of flight crew
TL3 Adverse System Interactions
TL4 System Failure Symptom
OL13 Mission Aborts
HL10 Recovery Decision Made by MCE Crew
HL8 Risk Perception of Flight Crew

Fig. 4. UAS operation risk CLD model in the crew resource and technical system levels

important indicator to evaluate a UAS fleet's operation training investment. Moreover, the contributors for the flight crew experience formation involve not only the formal personal training but also the mission learning behavior occurring in routine and emergency tasks, such phenomenon on personal knowledge and skill variation has been discussed extensively in job-training relevant researches [20]. It explains why the accumulation of flight crew experience always presented a delay process compared to the training investment.

(2) R_1 (EL_3-HL_2-HL_{11}-EL_9) and R_2 ($TL_{2.1}$-TL_3-$TL_{2.1}$)
Similar to the risk dynamics in the system technical level, the accident investigation (OL_6) can also enforce the revision of operation procedure (HL_{12}) with a delay process, but such procedure modification can also reduce the familiarity of flight crew on current tasks. The loop dominance between the B_1, B_2 and R_1 loops determines the actual experience level of flight crew with a dynamic view. The reinforcing loop R_2 illustrates such phenomenon. Meanwhile, this study uses a simplified model to describe the system failure interactions, such as the loss of electric power supply (e.g., engine failure) always induced the failure of data communication and initialize a lost link profile which always jeopardize the flight crew's routine operation, when the backup battery exhausted simultaneously.

(3) R_3 (HL_9-HL_{13}-EL_4-HL_9) and R_4 (HL_1-HL_8-HL_{11}-HL_{12}-HL_1)
This group models the effects of flight crew experience on the emergency level indicators of UAS operation process. This reinforcing loop R_1 describes how the safety

commitments formed under mission tempo and mission training affect the MCE crew's decision on recovery the UAS vehicle in advance when encountering system failure symptom. In fact, this is often called active safety acts but always affects UAS mission accomplishment. In contrast, the flight crew might also choose to ignore such failure single to implement scheduled tasks against the risk of catastrophic system failure when the mission tempo is high. What strategy they will adopt often depends on the dominance of their safety commitment in their mind. Meanwhile, the reinforcing loop R_2 explains the risk mechanism that when possessing known system failure mode, the positive risk perception of flight crews helps them to solve the emergency issues and maintain the operation risk under an acceptable level.

5 Conclusion

Regarding the operation safety of large UASs is an emergency property of the social-technical system which performs the UAS operation processes, this study introduces a conceptual modelling approach to describe the risk interactions involving the effects of flight crew errors on large UAS operation safety. Based on the proposed HFACS framework for human error identification and learning from the mishaps of previous UAS types especially USAF MQ-1 Predator fleet, the proposed causal loop diagrams integrate the risk factors involving organizational, fight crew source and technical system dimensions rather than identifying static accidental factors in textual way without considering the dynamic interaction and time sequences between those factors.

The proposed model make some important findings indicated in UAS accident investigations easy to understand: the operation procedure modification derived from UAS accident investigation may introduce adverse effects on flight crew experience and it is the reason why such hindsight may suppress safety benefits of the training investments. Moreover, due to the multiple delay links between flight crew employment requirement and average crew experience, the relationship to characterize their causation is non-linear, which influences the operation organization's decision-making on crew resource and contributes to the potential human error shaping context.

The analysis on proposed models emphasize that the variation trends of loop domination can explain the risk interactions of UAS operation. The operation risk depends both on the technical system failure and flight crew experience gained through job training and self re-learning in missions. Increasing scheduled mission sorties and durations are always the instinctive response of the organization to cope with the UAS mishaps, considering the low operation cost features of UAS compared to manned aircrafts. Such mission tempo can always affect the safety commitments of flight crews when they encounter emergency conditions in operation. Those conceptual models on flight crew error related risk interactions provide a baseline for future quantitative SD stock-flow simulation which can be developed to organizational risk assessment tools, when considering the evaluation of medium-long term benefits of potential safety policy and management decision in the widely spreading UAS operations.

Acknowledgments. This research was sponsored by the National Science Foundation of China (No. 61803263). The first author also thanks Professor Karen B Marais for helpful advice during the process of this research.

References

1. ICAO. Unmanned aircraft systems (UAV), 1st edn. ICAO, Canada (2011)
2. US DOD. Report to Congress on Future Unmanned Aircraft Systems Training, Operation, and Sustainability. Technical report, Under Secretary of Defense for Acquisition, Technology and Logistics, Department of Defense, Washington, DC (2012)
3. European Aviation Safety Agency (EASA). A-NPA No 16-2005 Policy for Unmanned Aerial Vehicle (UAV) Certification; European Aviation Safety Agency: Cologne, Germany (2005)
4. Li, W.J.: Unmanned Aerial Vehicle Operation Management. Beihang University Press, Beijing (2011). (in Chinese)
5. Ramalingam, K., Kalawsky, R., Noonan, C: Integration of unmanned aircraft system (UAS) in non-segregated airspace: a complex system of systems problem. In: Proceeding of the 2011 IEEE International Systems Conference, pp. 1–8. IEEE Press, New York (2011)
6. Schaefer, R.: Unmanned Aerial Vehicle Reliability Study, OSD UAV Reliability Study. Technical report, Office of the Secretary of Defense, Washington, DC (2003)
7. Nullmeryer, R.T., Herz, R., Montijo, G.A.: Training interventions to reduce air force predator mishaps. In: 15th International Symposium on Aviation Psychology, Dayton, OH (2009)
8. Nullmeryer, R.T., Herz, R., Montijo, G.A., Leonik, R.: Birds of prey: training solutions to human factors issues. In: 10th the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Dayton, OH (2007)
9. Wiegmann, D.A., Shappell, S.A.: A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System, pp. 45–56. Ashgate, Burlington (2003)
10. Williams, K.W.: A Summary of Unmanned Aircraft Accident/Incident Data: Human Factors Implications. Technical report, Civil Aerospace Medical Institute, FAA, Oklahoma City, OK (2004)
11. Murata, T.: Petri nets: properties, analysis and applications. Proc. IEEE **7**, 54–58 (1989)
12. Sterman, J.D.: Business Dynamics: Systems Thinking and Modeling for A Complex World. Irwin/Mac-Graw Hill, Boston (2000)
13. Lu, Y., Zhang, S.-G., Hao, L., Huangfu, H.-Y., Sheng, H.: System dynamics modeling of the safety evolution of blended-wing-body subscale demonstrator flight testing. Saf. Sci. **89**, 219–230 (2016)
14. Dulac, N., Owens, B., Leveson, N.G.: Demonstration of a new dynamic approach to risk analysis for NASA'S constellation program. MIT CSRL Final Report to the NASA ESMD Associate Administrator (2007)
15. Bouloiz, H., Garbolino, E., Tkiouat, M., Guarnieri, F.: A system dynamics model of behavioral analysis of safety conditions in a chemical storage unit. Saf. Sci. **58**(1), 32–40 (2013)
16. Wolstenholme, E.F.: Toward the definition and use of a core set of archetypal structures in system dynamics. Syst. Dyn. Rev. **19**(1), 7–26 (2003)
17. Marais, K.B., Saleh, J.H., Leveson, N.G.: Archetypes for organizational safety. Saf. Sci. **44** (7), 565–582 (2006)

18. Tvaryanas, A., Thompson, W., Constable, S.: Human factors in remotely piloted aircraft operations: HFACS analysis of 221 mishaps over 10 years. Aviat. Space Environ. Med. **77** (7), 724–732 (2006)
19. Lu, Y.: Feedback process-based UAV system accident analysis and system safety dynamics investigation. Ph.D. thesis. Beihang University, Beijing (2015). (in Chinese)
20. Lu, Y., Marais, K.B., Zhang, S.G.: Conceptual modeling of training and organizational risk dynamics. Proc. Eng. **80**, 313–328 (2014)

Checklist and Alert Language: Impact on ESL Pilot Performance in Airline Operations

Dujuan B. Sevillian(✉)

Human Factors and Transport Systems Engineering,
Cranfield University, Bedford, UK
d.sevillian@cranfield.ac.uk

Abstract. Flight crewmembers utilize checklists during typical phases of flight, which may also encompass non-normal conditions. Written English language on checklists combined with crew alerting can be used by flight crewmembers to read and comprehend system related issues, and respond to system conditions on the flight deck. Design and integration of English language on checklists and alerting systems should provide information that can be utilized for flight decision-making purposes. English language can be challenging for English as-a-second language flight crewmembers. Review of literature suggests that ESL adults experience fundamental challenges with reading and interpreting written English language text corpora based on their background knowledge, English language proficiency, and contextual use of written English language in airline operations. This paper provides a survey of ESL flight crew performance issues when they use checklists and alerting systems during non-normal conditions. Survey results indicated that flight crewmember use of written English language checklists has an impact on their performance in airline operations. Design and integration of written English language on checklists and alerting systems were factors leading to ESL flight crewmember procedural divergence and misunderstandings. Flight crewmembers' metacognitive strategy use, background knowledge, and their English language proficiency (reading comprehension level), were factors that impacted their performance and flight safety. Future studies should focus on ESL flight crewmember use of written English language on checklists and alerting systems and impact on flight crewmember performance in airline operations.

Keywords: Lexis · Human performance · System safety · Flight deck · Crew station design · Cognition

1 Introduction

English language can be considered challenging to read and interpret by adults in various sociotechnical environments. English as-a-second language (ESL) adult reading comprehension has the potential to be impacted by design and integration of written English language vocabulary words on alert and information systems. Adult English

Note: Literature Review, Methods, and Discussion sections are from Sevillian (2017).

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 284–307, 2019.
https://doi.org/10.1007/978-3-030-22507-0_23

language proficiency can also be a factor that influences their performance, which may impact ESL adult ability to read and understand word meaning. In the maritime industry, ESL seafarer's misunderstandings while they read and comprehended English language led to accidents (MAIB 2005). These accidents were related to seafarer's ability to read and understand vocabulary words on operational safety documentation. Aviation industry has indicated use of technical information by ESL crewmembers is also challenging, especially when using documentation in an operational environment. Drury and Ma (2003) found that maintenance personnel experience difficulties reading and comprehending safety information related to tasks. On the flight deck, use of written English language by ESL flight crewmembers has been noted as a factor influencing their performance. According to IAC (2013), ESL flight crewmember ability to adequately read and understand operational procedures and complex vocabulary words/sentence structure can negatively impact ESL flight crewmember competency when they read English language. After the airplane crash investigation of Tatarstan Airlines, the Interstate Aviation Committee (IAC) found that flight crewmember English language proficiency was a factor that influenced the crash. Particularly, the accident investigation team found that the Russian civil aviation authority did not levy requirements for Russian flight crewmembers to read and understand English language, with adequate proficiency levels (IAC 2015). The investigation also revealed that the International Civil Aviation Organization (ICAO) needed to update their English Language Proficiency Requirements (ELPRs) to include reading proficiency in English language. Current ICAO ELPRs address communication when using radiotelephony and ICAO phraseology. A previous aircraft accident in 2012 involving an ATR-72 VP-BYZ indicated a need to design and integrate written English language on operational manuals clearly (IAC 2013). The accident indicated that flight crewmember proficiency levels (ICAO ELPRs) are not the only requirements for proficiency. Flight crewmember reading comprehension of English language is a critical element that can impact their English language proficiency. Furthermore, IAC (2015) report indicated that English language vocabulary words and structure are ambiguous and can lead to misunderstandings. It was indicated that certain flight control procedures followed by flight crewmembers were misunderstood, which was a factor that led to the accident. Finally, the report concluded that flight crewmember English language proficiency was not adequate when they read flight manuals (e.g. Flight Crew Training Manuals), and their proficiency was found to be less than adequate during training.

Western built flight decks are designed to provide alerts and procedures that assist flight crewmembers with decision-making. Goal of designing alerts and procedures is to provide alert style and procedural guidance that corresponds to flight crewmember tasks on the flight deck (e.g. Quick Reference Handbook). Essentially, written English language should be designed and integrated with appropriate format, so that information can be used effectively by crewmembers to complete assigned tasks. According to Barshi et al. (2016), there are four aspects to consider when designing and integrating English language on checklists. These aspects are as follows: (1) consistent utilization of vocabulary words, (2) common word meaning, simple syntax (3) acronym and abbreviation use (4) appropriate vocabulary word use. Consistent utilization of vocabulary words is the process of using common wording on the flight deck. This provides crewmembers with ability to understand checklist information alongside flight deck terminology. Use of common word meaning provides flight crewmembers with ability to shape their mental model regarding vocabulary words used for particular tasks. Simple syntax can allow flight crewmembers to read and comprehend information in a timely manner. Acronyms and abbreviations are used often on crew alerting and information systems, but they should only be used when flight crewmembers are familiar with the terminology. For example KT is 'knots' and FL is 'flight level'. An example of an acronym is mode control panel 'MCP'. Confusion can occur between flight crewmembers if abbreviated forms of words and acronyms are used inappropriately. Use of appropriate vocabulary words is related to ensuring proper use of aviation English is standardized on alerts and operational documentation. In other words, information should be clear, concise, and provide the operator with the ability to make informed decisions. Words that are complex to read and understand may impact flight crewmembers' ability to respond to an alert action. Formatting written English language is also a factor that can influence flight crewmember performance on the flight deck. Barshi et al. (2016) indicated that conditional statements and implementation of warnings and cautions should be designed and integrated on checklists appropriately. Conditional statements often contain words that provide emphasis on actions that need to be completed. On the flight deck, non-normal procedures contain conditional statements that are used by flight crewmembers to determine crew actions needed to resolve issues related to system operations. Conditional statements should be structured in a format that is comprehensible for the user.

Considering previously discussed industry issues related to flight crewmember interaction with English language on the flight deck, what is the impact on ESL flight crewmember ability to read and comprehend written English language during non-normal conditions? What types of written English language impact ESL flight crewmember performance on the flight deck? What is the impact of flight crewmembers' English language proficiency on their ability to read and comprehend English language on the flight deck? What types of metacognitive strategies do flight crewmembers utilize while reading English language? These questions will be answered throughout literature review analyses and discussions, as well as throughout the researcher's study.

2 Literature Review

Research on human capabilities and limitations on the flight deck has provided the aviation/aerospace industry with an abundance of data, which has focused on ensuring the flight crewmembers have clear mental models on use of different types of information on the flight deck. As information on the flight deck is provided in different forms, it is important that design and integration of written English language on alerting and information systems (e.g. Electronic Centralized Aircraft Monitor (ECAM)/QRH) is consistent with flight crewmember expectations. Consistency in text corpora design, vocabulary word use, text genre, and sentence syntax are some factors that have the potential to impact ESL flight crewmembers reading comprehension performance. Following literature review provides an overview of factors that can lead to ESL adult misunderstandings when they read and comprehend English language. The review also provides and understanding of how misunderstandings can impact ESL adult performance in various sociotechnical environments.

According to Nielsen-Bohlman and Institute of Medicine (2004), ability to read and understand English language requires adults to have adequate knowledge, skills, and abilities (KSAs) when reading and comprehending English language. Adults should also have adequate proficiency when reading and comprehending English language. Adult proficiency in English language can lead to adequate reading comprehension performance. Yildiz-Genc (2009) indicated that ESL adults experience difficulties with reading and comprehending English language. Syntax, word meaning, and text genre, are just some of the factors that influence ESL adult ability to reading and comprehend English language. What processes or metacognitive strategies do ESL adults utilize to read and understand English language? Metacognition is operationally defined as the way in which an individual understands their cognitive processes. Metacognition helps individuals organize their thoughts/ideas to assess a situation or condition. A study conducted by Yildiz-Genc (2009) utilized 15 ESL adults with intermediate English language proficiency. No time constraint was levied on ESL adults and they used bottom-up and top down strategies to read and comprehend English language. Bottom-up strategy considers how an ESL adult may comprehend information considering a linear text flow. Decoding syntax a feature of bottom up strategy that can be used to decode information in a sentence. Adult English language proficiency and vocabulary knowledge is a factor that influences their ability to read and understand English language. Use of top down-strategy by ESL adults enables them to use previous knowledge to read and understand information in sentence syntax. Adults may use background knowledge of information to help them throughout the reading comprehension process. Results from Yildiz-Genc's (2009) study indicated that when ESL adults used bottom up strategies to read and comprehend English language, vocabulary word meaning challenged them, and they used previous sentences to interpret and connect their ideas to understand information they read. Adults also translated words, sentences, and phrases to understand sentence meaning. Furthermore, they re-read information to help them interpret information in sentences. Finally, top down processing was used their background knowledge to understand sentence meaning and vocabulary words. Hammadou (1991), Lin and Chern (2014) have also indicated that ESL adult use background knowledge understand information in sentences. A study conducted by Fatemi et al. (2014) focused on understanding the effects of ESL adult reading comprehension when they used top down and bottom up strategies. Eighty ESL adults were utilized for the study and each participant was proficient with written English language. The 80 adults were split into two groups (top down strategy/bottom up strategy cognitive styles). Results indicated that participants that used bottom up strategy performed better than participants using top down strategy. These results are likely due to the differences in cognitive reading style. Participants that used top down strategy did not comprehend text in the same way as participants using bottom up strategy. Participants that used bottom up strategy were accustomed to using decoding methods to critically analyze text versus participants that used top down strategy, which were accustomed to using their background knowledge to assess reading and comprehension of information.

Overall, Yildiz-Genc's (2009) and Fatemi et al. (2014) studies reveal that strategy use by ESL adults can be helpful when they read and comprehend English language. Depending on the type of strategy utilized, adults may perform differently based on their ability to read and comprehend information. Adult English language proficiency is a factor that influences type of strategy that adults may utilize. Previously discussed theories could be a potential influence on how ESL flight crewmembers perceive and process English language, through use of strategies. The type of strategy flight crewmembers use could potentially impact their ability to perform when responding to crew alerts and using QRH checklists. Next section provides an overview of how text genre influences adult understanding of information.

Text genre can be a factor that influences ESL adult ability to read and comprehend text/text corpora. Abdul-Hamid and Samuel (2012) studied the impact of scientific text (text related to specific subject matter) on adult reading comprehension. Participant English language proficiency levels were proficient or less than proficient. Overall goal of the study was to determine if reading difficulty was observed between participants when they read two different types of scientific texts. Participants had background knowledge of the texts they read, however there were text corpora that had a percentage of vocabulary words that had the potential to be unfamiliar to participants. First text contained 592 words and the other text contained 744 words. Academic words and scientific words were observed combined in each of the texts. Academic words can be more common in text and are part of the Academic Word List (AWL) rather than scientific words. Scientific text/technical text can be found in information that is specific to a particular industry (i.e. nuclear industry). Participants highlighted words they were unfamiliar with in the text they read. Omission of words was observed in the study as well as re-reading text for reading comprehension purposes. Results indicated that participants' proficiency level could have been a factor that led to their difficulties reading text. Park (2010) focused on a study that measured the effects of expository text (cause and effect) on ESL adult reading comprehension. The study contained 115 participants and they were studying English language for academic credit, with a focus on engineering and science. All participants had approximately 10 years of experience with using English language, and many of the participants had experience with English language in different regions of the globe such as United States of America. Many participants self rated themselves as having adequate knowledge of English language and some indicating somewhat adequate knowledge of English language. When participants self rate their English language proficiency it can provide details on how they interpret English language and challenges they may experience (Yeh and Genter 2005). Results from Park's (2010) study indicated participants had strong use of metacognitive strategies when they read expository text with a technical emphasis versus novel text. Participants highlighted text and re-read text for reading comprehension purposes. Rouhi et al. (2015) and Storch (2001) indicated that highlighting information in expository text is an indication that the ESL reader understands the structure (cause and effect). They also indicated that background knowledge in the subject is important when reading expository text. There was also a low-cohesion factor (explanations are less perceptible in the structure of text) in novel text rather than expository text.

Overall, Abdul-Hamid and Samuel (2012) and Park (2010) provide evidence that text genre can influence reading comprehension. Studies also revealed that when adults self rate themselves on their English language proficiency, this is an adequate indicator of their proficiency level. Adult experience with use of English language and metacognitive strategies are indicators that explain adult reading comprehension abilities. Previously discussed studies reveal the need to further research effects of ESL flight crewmembers use of crew alerting systems and QRH checklists during non-normal conditions. There is a potential that text genre could be different on QRH checklists, and flight crewmember ability to understand different types of text genre may influence their reading comprehension performance. For example, what is the impact to flight crewmembers that do not have adequate experience with use of English language on alert systems and QRH checklists? Does their proficiency level impact their ability to read and understand English language on alert systems and QRH checklists? These factors will be further discussed in the researcher's study.

You (2009) developed a study that focused on ESL adult ability to read and comprehend information on computer screens versus paper format. Two texts that were familiar and unfamiliar were utilized for the experiment design. Participant proficiency levels were low, medium, or high. Text length was 340 words and each of the readings was expository text genre. Results indicated that participants performed satisfactory. Participant background knowledge was better when they read English language from paper rather than computer screen. Participants were more accustomed to reading information on paper and using metacognitive strategies rather than on computer screen. Participants with medium and high proficiencies performed better reading text in the same format, rather than participants that read text in a different format. A study conducted by Park et al. (2014) focused on English language abbreviations. Seven participants from different regions of the globe had an English language proficiency of satisfactory. Two participants had technical background knowledge, while the other participants had academic/business knowledge. Participants had experience using English language in the United States and had knowledge of the text they read. Results indicated that acronyms were difficult to read and background knowledge was used to understand acronyms. Participants also utilized dictionary sources to understand the acronyms.

You (2009) and Park et al. (2014) studies indicate that text length and abbreviated text have an impact on how well ESL adults read and understand English language. Adult proficiency levels and technical background knowledge are factors that also influence how well adults read and interpret English language. Both authors indicated that use of metacognitive strategies by adults is influenced by level of English language proficiency. On the flight deck, ESL flight crewmembers use alert systems and checklists; therefore vocabulary words and checklist items should be adequately designed so they may be interpreted well by flight crewmembers. As many flight crewmembers may use background knowledge of English language from training or experience using English language, design and integration of information on checklists and alert systems must be written so they are understood from a variety of flight crewmembers with different linguistic backgrounds.

The literature review provided an overview of factors that influence ESL adult ability to read and understand written English language in socio-technical environments. Design and integration of English language has potential to impact adult performance. In particular, background knowledge of text is a factor that impacts adult ability to read and understand information. Adult proficiency level influences metacognitive strategy use and amount of metacognitive strategies utilized to read and interpret information. The type of words used in text corpora (i.e. academic words, technical words), influence adult reading and comprehension of information. On the flight deck, ESL flight crewmembers English language proficiency level, background knowledge, metacognitive strategy, variation of strategies utilized, and experience using English language, can influence how well flight crewmembers read and interpret information. It can also impact how they respond to non-normal conditions on the flight deck. The next sections provide an overview of the impact of ESL flight crewmember use of written English language on the flight deck.

3 Methods

A qualitative research study was conducted with 19 ESL flight crewmembers. Term flight crewmember is also known as roles captain/first officer. Each flight crewmember had experience flying large transport category aircrafts, such as the Embraer Regional Jet (ERJ). Flight crewmembers had Air Transport Pilot (ATP) ratings. All flight crewmembers had experienced with English language throughout their initial schooling (e.g. grade school) and secondary school—college education. For the purposes of this study, flight crewmembers' English language experience was considered background knowledge. The ICAO ELPRs level ratings were between four and six. Level four is considered operational use of English language and level six is more the satisfactory use of English language. Even though flight crewmember ICAO ELPRs level ratings are related to flight crewmember communication while using radiotelephony, the data was collected to understand influences that may impact flight crewmember background knowledge of English language. Flight crewmember reading comprehension levels were collected to understand how well they read and comprehend written English language. Each of the 19 flight crewmembers rated themselves on their general use of English language (command of English language in non-socio-technical environments), and proficiency when they read and comprehend written English language on alerts and the QRH on the flight deck (i.e. technical information on the flight deck). Flight crewmember proficiency levels were considered Reading Comprehension Levels (RCLs). Flight crewmember proficiency levels were either rated as low-intermediate (L-I), intermediate (I), or high-level (H). Low-intermediate English language proficiency indicated flight crewmember understanding of English language was adequate, but they had issues with sentence syntax and words. Flight crewmembers with intermediate-level proficiency indicated they required more knowledge of English language. Flight crewmembers with High-level English language proficiency indicated they were comfortable with reading and comprehending written English language. The following demographics were provided for the study (Tables 1, 2 and 3):

Table 1. Flight crewmember Demographics (N = 19)

Demographics	Pilot 1	Pilot 2	Pilot 3	Pilot 4	Pilot 5	Pilot 6	Pilot 7	Pilot 8
Country of origin	Ecuador	Ecuador	Ecuador	Ecuador	Brazil	Brazil	Ecuador	Trinidad
Age	53	32	43	29	34	50	37	51
Airline years of experience	15	8	11	4.5	10	6	10	8
Native language spoken	Spanish	Spanish	Spanish	Spanish	Portuguese	Portuguese	Spanish	Caribbean Dialect
English language learned/country	Grade School/ Ecuador	Grade School/ Ecuador	Grade School/ U.S.	University/ U.S.	University/ South America	University/ U.S.	University/ U.S.	University/ Trinidad
ICAO ELPR level	Level 6	Level 6	Level 6	Level 6	Level 4	Level 4	Level 5	Level 6
Self-rated English language RCL (General use of English language)	I-Level	I-level	I-Level	I-Level	I-Level	H-Level	H-Level	H-Level
Self-rated RCL: English language on crew alerting systems and QRH checklists	I-Level	I-Level	H-Level	L-I Level	L-I Level	H-Level	I-Level	H-Level

Table 2. Flight crewmember Demographics (N = 19)

Demographics	Pilot 9	Pilot 10	Pilot 11	Pilot 12	Pilot 13	Pilot 14	Pilot 15	Pilot 16
Country of origin	Brazil	Brazil	Brazil	Brazil	Jordan	Jordan	Jordan	Jordan
Age	36	28	45	41	32	25	38	28
Airline years of experience	12	6	17	11.5	3	2	13	3
Native language spoken	Spanish	Spanish	Spanish	Spanish	Arabic	Arabic	Arabic	Arabic
English language learned/country	Grade school/ secondary/ U.S.	Secondary school/ U.S.	Secondary school/ U.S.	Secondary school/ U.S.	Pre-school/ Jordan	Pre-school/ Jordan	Pre-school/ U.S.	Pre-school/ Jordan
ICAO ELPR Level	Level 4	Level 4	Level 4	Level 4	Level 5	Level 5	Level 6	Level 6
Self-rated English language RCL (General use of English language)	I-Level	I-Level	I-Level	H-Level	H-Level	H-Level	H-Level	H-Level
Self-rated RCL: English language on crew alerting systems and QRH checklists	L-I level	L-I Level	H-Level	L-I Level	H-Level	H-Level	H-Level	H-Level

Most flight crewmembers country of origin was Brazil. Second most frequent country of origin was Ecuador, followed by Jordan. Trinidad, United States of America (USA), Bulgaria, and Colombia were also flight crewmembers country of origin. Average age was 36 years old. Flight crewmembers most common spoken language was Spanish, Arabic, Portuguese, Caribbean dialect, and Bulgarian. The researcher led face-to-face interviews with 19 flight crewmembers. Data from interviews was

Table 3. Flight crewmember Demographics (N = 19)

Demographics	Pilot 17	Pilot 18	Pilot 19
Country of origin	Colombia	U.S.	Bulgaria
Age	22	26	37
Airline years of experience	4	1	4
Native language spoken	Spanish	Spanish	Bulgarian
English language learned/country	University/ U.S.	Pre-school/ U.S.	Pre-school/University as exchange student in U.S.
ICAO ELPR Level	Level 5	Level 5–6	Level 6
Self-rated English language RCL (General use of English language)	I-Level	I-Level	H-Level
Self-rated RCL: English language on crew alerting systems and QRH checklists	I-Level	I-Level	H-Level

recorded, coded, and themes were established based on the data. Researcher developed a questionnaire to collect data on flight crewmember performance when they read and comprehend information on alert systems and QRH checklists. Questionnaire focused on flight crewmember self rated reading comprehension proficiency levels, background knowledge of English language, and metacognitive strategies flight crewmembers utilized when they read and comprehend English language on alerting systems and QRH checklists. Follow-up discussions between the researcher and flight crewmembers were conducted. Researcher's coding method will be described in a future section.

4 Limitations

Information collected from surveys was generic to alerting systems and QRH checklists. The study did not measure flight crew performance, with respect to their ability to interpret vocabulary words and text genre, and measurement of workload when they read and comprehend written English language. These types of variables limited the scope of the researcher's study.

5 Coding Method

Researcher utilized a transcription template that consisted of coding information collected from interviews held between the researcher and flight crewmembers, and questionnaires that flight crewmembers completed. Coding schema was related to flight crewmember demographics, related to their ability to read and comprehend English language, background knowledge, English language proficiency (reading comprehension level), metacognitive strategies, crew alerting design/integration factors, and QRH checklist design/integration factors. Flight crewmember performance and flight safety related impacts were also coded.

6 Inter-rater Reliability

Researcher consulted two flight systems experts to review coding from interviews and questionnaires. Their background was in system safety and ESL flight crewmember performance. Experts used the previously discussed coding schema to determine if they could code information from the interviews and questionnaires and determine level of agreement. Results showed that there was substantial inter-rater reliability ($k = 1$).

7 Results

Results from the interviews and questionnaires indicated flight crewmembers noted several challenges with their ability to read and comprehend information on alert and information systems. High percentage of flight crewmembers indicated they use metacognitive strategies to read and interpret English language on alert systems and QRH checklists. Flight crewmembers noted that when they read and comprehend information on QRH checklists/alert systems together to solve system errors on the flight deck, their reading comprehension was negatively impacted. Flight crewmembers also indicated flight safety was impacted as a result of their ability to read and comprehend information on alerts and QRH checklists. Next results provide a review of flight crewmember self rated RCLs (proficiency levels), including their general use of English language and use of English language on alert systems and QRH checklists. Additionally, flight crewmember background knowledge factors, vocabulary words/text genre knowledge, metacognitive strategies use, and proficiency level results are provided (Table 4).

Table 4. English language proficiency factors

Description	Flight crewmembers percentage
Self rated English language proficiency RCL of general use of English language (L-I)	0/19 (0%)
Self rated English language proficiency RCL of general use of English language (I)	10/19 (∼53%)
Self rated English language proficiency RCL of general use of English language (HL)	9/19 (∼47%)
Self rated English language proficiency RCL of English language on crew alerting systems and QRH checklists (L-I)	5/19 (∼26%)
Self rated English language proficiency RCL of English language on crew alerting systems and QRH checklists (I)	5/19 (∼26%)
Self rated English language proficiency RCL of English language on crew alerting systems and QRH checklists (HL)	9/19 (∼47%)

Flight crewmembers had a variety of written English language proficiency levels with respect to their RCL of general English language, alerting systems and QRH checklists (Table 5).

Table 5. Flight crewmember Background knowledge factors

Description	Flight crewmembers percentages
English language-ICAO ELPR Level 4, 5, 6	19/19 (100%)
Preliminary School (Grade School) non-western region experience reading and speaking English language	3/19 (∼16%)
Preliminary School (Grade School) western region experience reading and speaking English language	7/19 (∼37%)
Secondary School (University) non-western region experience reading and speaking English language	0/19 (0%)
Secondary School (University) western region experience reading and speaking English language	9/19 (∼47%)
ATP Certification (ability to read English language)	19/19 (100%)
Airline years of experience using crew alerting systems and QRH checklists	19/19 (100%)

Flight crewmember ICAO proficiency levels were between 4–6. All flight crewmembers had an ATP certification and years of experience using alerting systems and QRH checklists. Flight crewmember English language experience was different with respect to institution type and western/non-western regions (Table 6).

Table 6. Flight crewmember Vocabulary Words/Text Genre Background knowledge factors

Description	Flight crewmembers percentage
Knowledge of English language text genre on crew alerting systems (e.g. technical text)	19/19 (100%)
Knowledge of English language text genre on QRH checklists (e.g. technical text)	19/19 (100%)
Knowledge of English language elements on QRH checklists (e.g. typographical elements)	19/19 (100%)
English language experience with conditional statements on QRH checklists (e.g. structure, noticing)	19/19 (100%)
Background knowledge of abbreviations/acronyms (e.g. short form and/or long form)	19/19 (100%)
Background knowledge of text format on crew alerting systems and QRH Checklists (e.g. authentic, elaborated, or short text)	19/19 (100%)
ATP certification (knowledge of crew alerting systems/QRH checklists)	19/19 (100%)
Background knowledge of vocabulary word type on crew alerting systems	19/19 (100%)
Background knowledge of vocabulary word type on QRH checklists	19/19 (100%)

Flight crewmembers had experience with vocabulary words and text genre background on alerting systems and QRH checklists (Tables 7 and 8).

Table 7. Flight crewmember metacognitive strategy use and proficiency level factors

Description	Flight crewmembers percentage	Flight crewmembers English language proficiency and percentage (RCL proficiency general English language)	Flight crewmembers English language proficiency and percentage (crew alerting systems and QRH checklists RCL proficiency)
Re-reading text	10/19 (∼53%)	∼32% I-level; ∼21% H-Level	∼21% I-level; ∼21% L-I level; 11% H-level
Paraphrasing text	0/19 (0%)	N/A	N/A
Underlining text	2/19 (∼11%)	∼11% I-Level	∼11% L-I Level
Referencing other Resources to clarify information (e.g. dictionary)	1/19 (∼5%)	∼5% H-Level	∼5% H-Level
Highlighting text	1/19 (∼5%)	∼5% I-Level	∼5% L-I Level
Translating written English language into ESL flight crewmembers native language	4/19 (∼21%)	∼5% I-Level; ∼16% H-Level	∼5% I-Level; ∼16% H-Level
Reverting back to native language to read English language	4/19 (∼21%)	∼21% I-level	∼5% I-Level; ∼16% H-Level
Reading aloud text on flight deck	2/19 (∼11%)	∼5% I-Level; ∼5% H-Level	∼11% H-Level

Table 8. Flight crewmember metacognitive strategy use and proficiency level factors continued

Description	Flight crewmembers percentage	Flight crewmembers English language proficiency and percentage (RCL proficiency general English language)	Flight crewmembers English language proficiency and percentage (crew alerting systems and QRH checklists RCL proficiency)
Monitoring reading comprehension	1/19 (∼5%)	∼5% H-level	∼5% H-level
Taking notes	2/19 (∼11%)	∼11% I-level	∼11% L-I level
Breaking apart sentences	3/19 (∼16%)	∼11% I-level; ∼5% H-level	∼5% L-I level; ∼11% H-level
Bottom up strategy (decoding text)	3/19 (∼16%)	∼5% I-level; ∼11% H-level	∼5% I-level; 11% H-level

(*continued*)

Table 8. (*continued*)

Description	Flight crewmembers percentage	Flight crewmembers English language proficiency and percentage (RCL proficiency general English language)	Flight crewmembers English language proficiency and percentage (crew alerting systems and QRH checklists RCL proficiency)

Top down strategy (prior knowledge of text; activating text schema)	5/19 (~26%)	~11% I-level;
		~16% H-level
			~5% L-I level;
			~5% I-level;
			~16% H-level

Interactive strategy (combination of bottom up and top down strategy use)	4/19 (~21%)	~5% I-level; ~16% H-level	~5% I-level; ~16% H-level
Monitoring reading speed	2/19 (~11%)	~5% I-level; ~5% H-level	~5% H-level; ~5% H-level
Skipping words/ omission of words	2/19 (~11%)	~11% I-level	~5% H-level; ~5% L-I level

Flight crewmembers utilize different metacognitive strategies to read and comprehend written English language. Flight crewmember metacognitive strategy use and English language proficiency levels were different when they read and interpret written English language (Tables 9 and 10).

Table 9. Crew alerting system design and integration factors as indicated by flight crewmembers

Description	Flight crewmembers Percentage	Flight crewmembers English language proficiency and percentage (RCL proficiency general English language)	Flight crewmembers English language proficiency and percentage (crew alerting systems and QRH checklists RCL proficiency)
Sentence length (short)	0/19 (0%)	N/A	N/A
Acronyms/abbreviations	6/19 (~32%)	~16% I-level; ~16% H-level	~16% I-level; ~16% H-level

Text genre (e.g. technical)	9/19 (~47%)	~32% I-level;
		~16% H-level
			~21% I-level;
			~21% H-level;
			~5% L-I level

Number of tokens in text	0/19 (0%)	N/A	N/A

Table 10. Crew alerting system design and integration factors as indicated by flight crewmembers continued

Description	Flight crewmembers percentage	Flight crewmembers English language proficiency and percentage (RCL proficiency general English language)	Flight crewmembers English language proficiency and percentage (crew alerting systems and QRH checklists RCL proficiency)
Authentic text	9/19 (47%)	~32% I-level; ~16% H-level	~21% I-level; ~21% H-level; ~5% L-I level
Sentence length (long)	1/19 (~5%)	~5% I-level	~5% I-level
Simplification of text	1/19 (~5%)	~5% I-level	~5% H-level
Vocabulary words type	5/19 (~26%)	~26% I-level	~11% I-level; ~11% H-level; ~5% L-I level

Flight crewmembers indicated several different written English language design and integration factors influenced their ability to read and interpret information on alerting systems. Flight crewmember English language proficiency level indicated differences with respect to English language design and integration factors that negatively impacted flight crewmember reading comprehension of English language on crew alerting systems (Table 11).

Table 11. QRH checklist design and integration factors as indicated by flight crewmembers

Description	Flight crewmembers Percentage	Flight crewmembers English language proficiency and percentage (RCL proficiency general English language)	Flight crewmembers English language proficiency and percentage (crew alerting systems and QRH checklists RCL proficiency)

Conditional statements	3/19 (~16%)	~11% I-level;
		~5% H-level
			~5% L-I-level;
			~5% I-level;
			~5% H-level

Number of token in text	3/19 (~16%)	~11% I-level; ~5% H-level	~11% L-I level; ~5% H-level

Authentic text	17/19 (~89%)	~47% I-level;
		~42% H-level
			~26% I-level;
			~26% L-I level;
			~37% H-level

(*continued*)

Table 11. (*continued*)

Description	Flight crewmembers Percentage	Flight crewmembers English language proficiency and percentage (RCL proficiency general English language)	Flight crewmembers English language proficiency and percentage (crew alerting systems and QRH checklists RCL proficiency)

Sentence length (long)	5/19 (~26%)	~16% I-level;
		~11% H-level
			~11% L-I level;
			~11% H-level;
			~5% I-level

Simplification of text	1/19 (~5%)	~5% H-level	~5% H-level
Acronyms/abbreviations	4/19 (~21%)	~11% I-level; ~11% H-level	~16% I-level; ~5% H-level

Text genre (e.g. technical)	17/19 (~89%)	~47% I-level;
		~42% H-level
			~26% I-level;
			~26% L-I level;
			~37% H-level

Vocabulary words type	14/19 (~74%)	~42% I-level;
		~31% H-level
			~16% I-level;
			~26% L-I level;
			~31% H-level

Sentence length (short)	0/19 (0%)	N/A	N/A

Flight crewmembers indicated several different written English language design and integration factors that impacted their ability to read and interpret information on QRH checklists. Flight crewmember English language proficiency levels indicated differences with respect to English language design and integration factors that negatively impacted their reading comprehension of English language on QRH checklists (Table 12).

Table 12. Flight safety impact factors as indicated by flight crewmembers

Main theme: ESI flight crewmembers flight safety impact	Percentages	Flight crewmembers English language proficiency and percentage (RCL proficiency general English language)	Flight crewmembers English language proficiency and percentage (crew alerting systems and QRH checklists RCL proficiency)
Improper system diagnosis - Difficulty understanding abbreviations and acronyms	1/19 = ~5%	~5% I-level	~5% I-level

(*continued*)

Table 12. (*continued*)

Main theme: ESI flight crewmembers flight safety impact	Percentages	Flight crewmembers English language proficiency and percentage (RCL proficiency general English language)	Flight crewmembers English language proficiency and percentage (crew alerting systems and QRH checklists RCL proficiency)

Long processing time of information
-Due to translation of words into native language, highlighting/underlining words on checklists
- Due to decoding abbreviations
	10/19 = ~52%	~31% I-level;
		~21% H-level
			~5% I-level;
			21% L-I level;
			~26% H-level

Workload impact - Very detailed QRH checklists - Challenging vocabulary words	5/19 = ~26%	~16% H-level; ~11% I-level	~21% H-level; ~5% L-I level
Frustration - Very detailed QRH checklists - Unknown words	1/19 = 5%	~5% H-level	~5% L-I level
Omission and misinterpretation of information - Skipping words due to misunderstanding - Reverting back to native language	2/19 = ~11%	~5% I-level; ~5% H-level	~5% H-level; ~5% I-level

Regarding flight crewmember English language proficiency levels, each flight crewmember indicated different proficiency levels with respect to their performance factors that negatively impacted flight safety.

8 Discussion

Previous literature indicated that ESL adult background knowledge of English language, knowledge of text genre/vocabulary words, and English language proficiency are key components to understand how well adults may read and comprehend written English language. The researcher's study indicated that all flight crewmembers had background knowledge of English language. They received English language instruction from a variety of educational institutional learning systems (e.g. university education). Many flight crewmembers had western region experience with English language (grade school and university) Flight crewmembers also had airline years of experience using written English language on crew alerting systems and QRH checklists. Therefore, flight crewmembers had background of vocabulary words/text genre background. Flight crewmembers' ATP ratings were utilized, as it was an indication they were able to read English language on the flight deck. As ECFR (2016) indicated, ATP rating is common for ESL airline flight crewmembers and is an indication that flight crewmembers must be able to read English language. The ICAO level of English language proficiency data collected indicated that all flight crewmembers met minimum requirements for ELPRs and some exceeded the requirements (ICAO 2004). Although flight crewmember ICAO ELPR levels were level four, five, and six, these levels do not provide an indication of how well flight crewmembers read and comprehend written English language. The IAC (2013) indicated that ESL flight crewmember ICAO ELPRs are not enough to assess how well flight crewmembers read and comprehend written English language. Therefore, self-rated English language proficiency levels were utilized and indicated each flight crewmember had different English language proficiency RCL with respect to their general English language reading comprehension. Additionally, flight crewmembers had dissimilar English language proficiency RCL reading and comprehending written English language on crew alerting systems and QRH checklists. Recall, utilization of ESL adult self-proficiency ratings are important, as they provide indicators of adults metacognitive strategy use, and how well they read and comprehend written English language on technical information, especially expository and instructional texts (Park 2010; Yeh and Genter 2005). Technical information was noted as challenging to many flight crewmembers regardless of the metacognitive strategy they utilized to read and understand written English language. Their use of metacognitive strategies to read and comprehend written English language on crew alerting systems and QRH checklists were different, and proficiency levels (general English language, crew alerting systems and QRH checklists) varied based on use of either crew alerting systems and/or QRH checklists. Regarding metacognitive strategy use by flight crewmembers, strategies utilized on QRH checklists (paper format) were different than crew alerting systems (displayed format). As Holder (2003) indicated, flight crewmember English language proficiency has the potential to be different based on their use of each of these systems (i.e. crew alerting systems and QRH checklists). Collectively, flight crewmembers' English language proficiency influenced their ability to read and comprehend written English language. Flight crewmembers had various English language proficiency levels, and each flight crewmember proficiency level influenced their ability to read information on crew alerting systems and QRH checklists. Altogether, aforementioned aspects were fundamental requirements needed to assess how well flight crewmembers read and understand written English language on crew alerting systems and QRH checklists, and challenges they experienced reading technical information. Next sections provide detailed discussions on the researcher's study.

As Smith-Jackson (2006) and Riley et al. (2006) indicated, understanding differences in flight crewmember cognitive processing of written English language is important, especially factors that may impact their performance. Written English language on crew alerting systems and QRH checklists should be evaluated, with respect to flight crewmembers cognitive ability to read and understand written English language on each of the systems (Burian 2006 and Holzinger et al. 2011). With respect to metacognitive strategies use by flight crewmembers, the researcher's study indicated differences in type of strategy utilized, number of metacognitive strategies utilized, and most common/least common strategy utilized to read and comprehend written English language. Most flight crewmembers utilized at least one metacognitive strategy to read and understand written English language, and there were many flight crewmembers with RCL proficiency H-level (general English language, crew alerting systems and QRH checklists) that utilized many metacognitive strategies to read and understand written English language on crew alerting systems and QRH checklists. As Park's (2010) study indicated, high self-rated proficiency ESL adults utilize more metacognitive strategies. On the other hand, in the researcher's study flight crewmembers with RCL proficiency I-level (general English language, crew alerting systems and QRH checklists) also utilized many metacognitive strategies. It was indicated that flight crewmembers with RCL proficiency I-level were also comfortable with using strategies to read written English language. Flight crewmembers with RCL proficiency L-I level indicated they utilized strategies to help guide them through the reading comprehension process. Anderson (2004) indicated that ESL adults read and interpret written English language utilizing mental models. In the researcher's study, flight crewmembers (sixteen percent) utilization of bottom up strategy (decoding text) was found. As Liu (2014) indicated, use of this model is dependent on ESL adult English language proficiency. Likewise, flight crewmembers (eleven percent) with RCL proficiency H-level and five percent with RCL proficiency I-level (general English language) use bottom up strategy (decoding text), while flight crewmembers (eleven percent) with RCL proficiency H-level and flight crewmembers (five percent) with RCL proficiency I-level (crew alerting systems, QRH checklists) utilize bottom up strategy (decoding text). It was indicated that flight crewmembers with RCL proficiency H-level had background knowledge of decoding words on crew alerting systems and QRH checklists. Additionally, flight crewmembers with RCL of H-level proficiency indicated they were comfortable using this strategy to read and understand written English language on crew alerting systems and QRH checklists. Use of top down strategy (background knowledge) by twenty-six percent of flight crewmembers was utilized more than bottom up strategy to activate their background knowledge/content schema of written English language text, on crew alerting systems and QRH checklists. Use of background knowledge by ESL adults to read and interpret English language is typical as indicated by Lin and Chern (2014), Hammadou (1991). In the researcher's study, flight crewmembers indicated they utilized English language skills they learned from their airline as mechanisms to read and understand written English language on crew alerting systems and QRH checklists. They considered their years of experience as an indicator of background knowledge of English language as well as the different types of checklists containing different layouts of technical information. Comparable to the flight crewmembers with RCL proficiency H-level that utilized bottom up strategy to read and understand written English language, flight crewmembers with RCL proficiency H-level also utilize top down strategy more than flight crewmembers with RCL proficiency I-level and L-I level. Flight crewmembers (sixteen percent) with RCL proficiency H-level utilize top down strategy, while eleven percent of flight crewmembers with RCL proficiency I-level (general English language) use top down strategy. On the other hand, flight crewmembers (sixteen percent) with RCL proficiency H-level, five percent I-level, and five percent L-I level (crew alerting systems, QRH checklists) use top down strategy. Flight crewmembers with RCL proficiency H-level indicated they were comfortable with written English language on crew alerting systems and QRH checklists because they were able to utilize their background knowledge of the systems. This finding is consistent with Yildiz-Genc's (2009) and You's (2009) study which indicated that background knowledge and familiarity with written English language indicates that ESL adults will read and understand written English language better than text that is unfamiliar to them. Twenty-one percent of flight crewmembers' indicated they use interactive strategy. Flight crewmembers (sixteen percent) were RCL proficiency H-Level and five percent were I-level (general English language), while flight crewmembers (sixteen percent) with RCL proficiency H-level and five percent I-level (crew alerting systems, QRH checklists) use interactive strategy. Flight crewmembers indicated that use of this strategy was due to their ability to decode and use background knowledge on sections of the QRH checklists. This finding is consistent with Fatemi et al.'s (2014) study. Flight crewmembers also indicated that familiarity with checklists items helped them recognize certain pieces of text. Re-reading text on crew alerting systems and QRH checklists was considered a strategy utilized by most flight crewmembers (fifty-three percent). Flight crewmembers (thirty-two percent) with RCL proficiency I-level and twenty-one percent of flight crewmembers with RCL proficiency H-level (general English language) utilized re-reading text strategy. Twenty-one percent of flight crewmembers that were RCL proficiency L-I level and twenty-one percent that were I-level use re-reading text strategy, while eleven percent of flight crewmembers with RCL proficiency H-level (crew alerting systems, QRH checklists) use re-reading text strategy. Flight crewmembers with RCL proficiency H-level indicated they only re-read text, if they did not understand information on checklists. On the other hand, flight crewmembers with RCL proficiency level I-level and L-I level indicated they re-read information to have a clearer picture of the system issue. In other words, flight crewmembers with RCL proficiency I-level and L-I level re-read checklist information as a practice to ensure they understood information, whereas, flight crewmembers with RCL proficiency H-level, only re-read information if they misinterpreted a word or sentence on a checklist. Flight crewmembers with RCL proficiency H-level indicated that sometimes very detailed checklists require certain words to be re-evaluated/re-interpreted. As Yildiz-Genc (2009) indicated, intermediate level ESL adults were more inclined to re-read sentences to understand the meaning. In the researcher's preliminary study flight crewmembers with RCL I-level indicated they re-read information as a common practice, not just to understand word or sentence meaning. Twenty-one percent of flight crewmembers' translate written English language on QRH checklists into their native language. Sixteen percent of flight crewmembers had RCL proficiency H-level and five percent I-level (general English language), while sixteen percent of flight crewmembers with H-level and five percent I-level (crew alerting systems, QRH checklists) translate written English language on QRH checklists back into their native language. As Hutchins et al. (2006, p. 5) indicated, "certain words may not be translated adequately and could destroy word meaning". In the researcher's study, long processing time of information was due to translation of checklists words and sentences into their native language. As Abdul-Hamid and Samuel (2012) indicated, translation of written English language into their native language led to ESL adults re-reading sentences. This was not the case in the researcher's study, rather flight crewmembers' reading time was long due to processing translated written English language words into their native language. They indicated they utilize translation strategy because their airline uses the strategy often to understand written English language on crew alerting systems and QRH checklists. Interestingly, ESL adult proficiency levels in Abdul-Hamid and Samuel (2012) study were either proficient or less than proficient. In the researcher's study, flight crewmembers' RCL proficiency was H-level or I-level, there were no flight crewmembers that translated written English language text, with RCL proficiency of L-I level. Therefore, the researcher's finding does not support this aspect of Abdul-Hamid and Samuel (2012) study, which indicated that less than proficient adults were negatively impacted by translation process. Twenty-one percent of flight crewmembers indicated they use reversion back to their native language strategy to understand written English language on crew alerting systems. Twenty-one percent of flight crewmembers with RCL proficiency I-level (general English language) indicated they use reversion strategy, while sixteen percent of flight crewmembers with RCL proficiency H-level and five percent I-level (crew alerting systems and QRH checklists) use reversion strategy. Flight crewmembers indicated they use this strategy as a common practice at their airline. As Kobayashi and Rinnert (1992) indicated, reverting back to English language can occur because ESL adult lack of understanding translated syntax meaning. This can result in inappropriate translation of technical information back into their native language. In the researcher's study, flight crewmembers' indicated they utilized this strategy because some aviation abbreviations and words are the same definition and are written fairly the same. Familiarity with words in their native language helps them as they process words on crew alerting systems when they use reversion strategy. As Larsen and Hansen (2010) indicated abbreviations and acronyms that are found in certain genres of text aid ESL adults with understanding their meaning due to their familiarity with the text. Additionally, this strategy did not lead flight crewmembers to incorrect translation of words into their native language. Referencing other resources to help clarify information (e.g. dictionary) was a strategy utilized by five percent of flight crewmembers. A flight crewmember with RCL proficiency H-level (general English language, crew alerting systems and QRH checklists) uses referencing other resources strategy to read written English language on crew alerting systems and QRH checklists. Five percent of flight crewmembers' use highlighting text strategy on QRH checklists. The flight crewmember had RCL proficiency I-level (general English language) and L-I level (crew alerting systems and QRH checklists). Flight crewmembers' (eleven percent) utilize taking notes strategy. Eleven percent of flight crewmembers' proficiency levels were RCL proficiency I-level (general English language) and L-I level (crew alerting systems and QRH checklists). According to Park's (2010) study, there were many ESL adults that utilized referencing and highlighting strategies to read and comprehend written English language text. In Park's (2010) study, note taking was the least utilized strategy. Additionally, Park's (2010) study indicated that more ESL adults had fairly good or not adequate English language proficiency, than high English language proficiency level ESL adults (English speaking and reading comprehension abilities). Contrary to Park's (2010) study, the researcher's preliminary study indicated that referencing and highlighting strategies were utilized the least by flight crewmembers with RCL of H-level, I-level, and L-I level (general English language, crew alerting systems and QRH checklists). Note taking strategy was not utilized the least by flight crewmembers, it was utilized more than referencing and highlighting text to read and interpret written English language on checklists. They indicated note taking helped them remember words they may see again on QRH checklists. Whereas, referencing and highlighting were indicated as a strategy utilized to access information on the checklists when they had a system malfunction/failure in an aircraft they flew. Monitoring reading comprehension was utilized by five percent of flight crewmembers. A flight crewmember with RCL H-level (general use of English language, crew alerting systems and QRH checklists) indicated use of monitoring reading comprehension strategy. Whereas, monitoring reading speed was commonly utilized by eleven percent of flight crewmembers. A flight crewmember with RCL proficiency I-level and a flight crewmember with H-level (general English language) use monitoring reading speed strategy. Both flight crewmembers indicated their RCL proficiency levels were H-level (crew alerting systems and QRH checklists). As Park's et al. (2014) study revealed, ESL adults with very good English language proficiency utilized monitoring reading comprehension to read and comprehend written English language. Part of Park's et al. (2014) study was corroborated in the researcher's preliminary study. One flight crewmember with high English language proficiency utilized monitoring reading comprehension to read written English language on QRH checklists. It was indicated that this was a practice the flight crewmember utilized to help set his expectations on the type of information he was about to read. Monitoring reading speed strategy was not indicated in Park's et al. (2014) study, but was utilized as a strategy by two flight crewmembers with high and intermediate level of English language proficiency in the researcher's preliminary study. Eleven percent of flight crewmembers' used skipping/omission of words on crew alerting systems and QRH checklists. Each flight crewmember (eleven percent) had RCL proficiency I-level (general English language), while eleven percent of flight crewmembers had RCL proficiency H-level and L-I level (crew alerting systems, QRH checklists). Each flight crewmember indicated they utilized skipping and omission of words if they did not understand written English language text. As Dordick (1996) indicated omission of words is due to ESL adults misunderstanding words, or unfamiliar words in text. As this was the case in the researcher's study, this strategy was also utilized by flight crewmembers with different levels of English language proficiency. As Abdul-Hamid and Samuel (2012) study revealed, ESL adults that were proficient with English language and less than proficient utilize skipping/omission strategy to understand written English language. Sixteen percent of flight crewmembers that utilize breaking apart sentences had a variety of RCL proficiency levels. Eleven percent of flight crewmembers with RCL proficiency I-level and five percent H-level (general English language) use breaking apart sentences strategy. On the other hand, five percent of flight crewmembers with RCL proficiency L-I level and eleven percent of flight crewmembers with RCL proficiency H-level (crew alerting systems, QRH checklists) indicated they utilized breaking apart sentences strategy. It was indicated that they use this strategy if they were unfamiliar with text or text seemed to be longer than expected on QRH checklists. Part of this finding is corroborated in Anderson (2003) study. In Anderson's (2003) study, it was indicated that intermediate level ESL adults utilized breaking apart sentences to understand written English language text. The researcher's study revealed that flight crewmembers with RCL proficiency H-level, L-I level, and I-level utilized breaking apart sentences to read and understand text on QRH checklists. Flight crewmembers (eleven percent) utilize underlining text on QRH checklists and had RCL proficiency of I-level and L-I level (general English language, crew alerting systems and QRH checklists). Flight crewmembers' indicated they utilized underlining strategy if they were unfamiliar with text, and if time permitted would go back and review the meaning of the word during a period of time that was not congested with other tasks. They also indicated they underlined text if it was unfamiliar to them in their native language. This finding is different from Rouhi et al. (2015) and Storch (2001) studies. They suggested highlighting text or providing emphasis to text is an indication that ESL adults were familiar with the structure of text. As flight crewmembers had background knowledge of text structure on QRH checklists, it is peculiar as to why they underlined text for a different reason than how Rouhi et al. (2015) and Storch (2001) studies explained use of this metacognitive strategy. Finally, eleven percent of flight crewmembers with RCL I-level and H-level (general English language) utilized reading aloud strategy. The flight crewmembers (eleven percent) also indicated they had an RCL proficiency of H-level (crew alerting systems, QRH checklists). Flight crewmembers' indicated they read aloud QRH checklists procedures and information on crew alerting systems, as this was a common practice at their airline. They also indicated use of this strategy to ensure that understood the QRH checklist procedure. As KNKT (2015) indicated, it is a common practice to read aloud procedures to understand information on crew alerting systems and QRH checklists.

9 Conclusion

Written English language factors on each of the systems previously discussed negatively impact flight crewmember performance. Flight crewmember English language background knowledge, text genre knowledge, and vocabulary words on crew alerting systems and QRH checklists, provide an understanding flight crewmembers familiarity with their use of English language. Flight crewmember English language proficiency is a factor that can impact flight crewmember ability to read and comprehend English language. Flight crewmember English language proficiency levels are essential for understanding their metacognitive strategy use to read written English language. Strategy type and amount of strategies utilized by flight crewmembers when they read and comprehend information on alert systems/QRH checklists is important. Flight crewmember use of strategies helps to understand how they interact between information on alert systems and QRH checklists. The ICAO English language proficiency levels provided by flight crewmembers did not match their reading comprehension level of general English lexis or their reading comprehension proficiency when interacting with alerting systems and QRH checklists. The ICAO English language proficiency levels should not be the only approach to achieve English language proficiency levels. Self-rating reading comprehension proficiency levels are an ample method of collecting data related to flight crewmembers' ability to read and comprehend English language.

Acknowledgments. Dr. Sevillian is a Sr. Scientist and Lead System Safety Engineer for the Boeing Company. His experience is within the confines of human systems integration and auto-flight systems design. "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". The researcher would also like to thank numerous airline pilot associations, flight crew, and subject matter experts that have assisted with development of this research study and future studies.

References

Abdul-Hamid, S., Samuel, M.: Reading scientific texts: some challenges faced by EFL readers. Int. J. Soc. Sci. Hum. **2**(6), 509 (2012)

Anderson, N.: Metacognitive reading strategy awareness of ESL and EFL learners. CATESOL J. **16**(1), 1–17 (2004)

Anderson, N.J.: Teaching reading. In: Nunan, D. (ed.) Practical English Language Teaching, pp. 67–86. McGraw-Hill, New York (2003)

Barshi, I., Mauro, R., Degani, A., Loukopoulou, L.: Designing Flightdeck Procedures. National Aeronautics & Space Adminstration (NASA)/TM 2016-219421 (2016)

Burian, B.K.: Design guidance for emergency and abnormal checklists in aviation. In: Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, San Francisco (2006)

Dordick, M.: Testing for a hierarchy of the communicative interference value of ESL errors. System **24**, 299–308 (1996)

Drury, C.G., Ma, J.: Do language barriers result in aviation maintenance errors? In: Human Factors and Ergonomics Society 47th Annual Meeting Proceedings, Denver, Colorado, October 13017 (2003)

ECFR: Electronic Code of Regulations. 14 CFR 61.153 subpart B (2016). http://www.ecfr.gov/cgi-bin/ECFR?page=browse. Accessed 27 Jan 2016

Fatemi, A.H., Vahedi, V.S., Seyyedrezaie, Z.S.: The effects of top-down/bottom-up processing and field-dependent/field-independent cognitive style on Iranian EFL learners' reading comprehension. Theory Pract. Lang. Stud. **4**(4), 686–693 (2014)

Hammadou, J.: Interrelationships among prior knowledge, inference and language proficiency in foreign language reading. Mod. Lang. J. **75**, 27–38 (1991)

Holder, B.E.: Improving the Boeing Quick Reference Handbook. NASA, Washington, DC (2003)

Holzinger, A., Baernthaler, M., Pammer, W., Katz, H., Bjelic-Radisic, V., Ziefle, M.: Investigating paper vs. screen in real-life hospital workflows: performance contradicts perceived superiority of paper in the user experience. Int. J. Hum.-Comput. Stud. **69**, 563–570 (2011)

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)

ICAO: Manual on the Implementation of ICAO Language Proficiency Requirements First Edition, Doc 9835 AN/453; ICAO: Washington, DC (2004)

Interstate Aviation Committee Air Accident Investigation Commission (IAC): Final Report on ATR-72-201 aeroplane Aircraft Accident; VJP-BYZ, Bermudas (2013)

Interstate Aviation Committee Air Accident Investigation Commission (IAC): Final Report on 737-500(53A) Tatarstan Airlines Aircraft Accident. VQ-BBN (Bermuda) (2015)

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)

Larsen, T.J., Hansen, R.A.: L2 processing and comprehension of complex syntactic structures in English medical texts. Student Masters thesis: Copenhagen Business School. Department of international studies and linguistics (2010)

Lin, Y.H., Chern, C.L.: The effects of background knowledge and L2 reading proficiency on taiwanese university students' summarization performance. Contemp. Educ. Res. Q. **22**(4), 149–186 (2014)

Liu, S.: L2 reading comprehension: exclusively L2 competence or different competences? J. Lang. Teach. Res. **5**(5), 1085–1091 (2014)

Maritime Accident Investigation Board-MAIB: Report on the investigation of the grounding of Jackie Moon; Dunoon Breakwater; Firth of Clyde, Scotland 1 September 2004: Report Number 5/2005 (2005)

Nielsen-Bohlman, L., Institute of Medicine (U.S.): Health literacy: A prescription to end confusion. National Academies Press, Washington, D.C (2004)

Park, Y.: Korean EFL college students' reading strategy use to comprehend authentic expository/technical texts in English. Student Dissertation: University of Kansas 2010-04-19 (2010)

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)

Riley, D.M., Newby, C.A., Leal-Almeraz, T.O.: Incorporating ethnographic methods in multidisciplinary approaches to risk assessment and communication: cultural and religious use of mercury in Latino and Caribbean communities. Risk Anal. **26**(5), 1205–1221 (2006)

Rouhi, A., Jafarigohar, M., Alavi, M.S., Asgarabadi, H.Y.: Task difficulty of macro-genres and reading strategies and reading comprehension. Int. J. Asian Soc. Sci. IJASS **5**(11), 656–677 (2015)

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.: Culture and warnings. In: Wogalter, M.S. (ed.) Handbook of Warnings, chap. 27, pp. 363–372. Lawrence Erlbaum Associates, Mahwah (CRC Press, Boca Raton) (2006)

Storch, N.: An investigation into the nature of pair work in an ESL classroom and its effect on grammatical development. Ph.D. thesis, Department of Linguistics and Applied Linguistics, The University of Melbourne (2001)

Yeh, D., Gentner, D.: Reasoning counterfactually in Chinese: picking up the pieces. In Proceedings of the Twenty-seventh Annual Meeting of the Cognitive Science Society, pp. 2410–2415 (2005)

Yildiz-Genc, Z.: An investigation on strategies of reading in first and second languages. Selected papers from 18th ISTAL, pp. 407–415 (2009)

You, C.: A comparative study of second language reading comprehension from paper and computer screen (2009)

Fixation Adjustment During the Landing Process and Its Relationship with Pilot Expertise and Landing Performance

Yanjin Sun[1,2,3], Jingyu Zhang[1,2](✉), Han Qiao[1,2],
Xianghong Sun[1,2](✉), Ping Qian[3], and Yang Song[3]

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology,
Beijing, China
{zhangjingyu, sunxh}@psych.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] China Eastern Airlines, Shanghai, China

Abstract. In order to examine whether there is a fixation adjustment and whether expert and novice pilots differ in their adjustment patterns during the landing process, and to identify specific fixation indicators that might be linked with landing performance, the eye movement data marked at different height (50–40 ft, 40–30 ft, etc.) of 29 captains and 28 copilots which were all serving in civil airlines with valid commercial license was collected by glasses eye-tractor during landing missions on Airbus320 full-motion simulators. The results show that: The landing performance of the experts was higher than that of the novices, experts had significantly better scores in terms a significant difference on the score of landing position, landing load, and handling stability, the difference between experts and novices on attitude, yaw angle and bank angle of the aircraft were only approaching significant. The study also found evidence suggesting the pilots do adjust their gaze points during the landing process and the experts made such an adjustment in a more salient manner. The pilots reduced their times of watch to the front part of the runway but increased their times of watching the rear part of the runway during the landing process. As compared to the novices, experts put much more attention to the front part of the runway from very beginning to almost halfway (50–30 ft height rang feet).

Moreover, earlier fixations on the rear part of the runway may result in deteriorated landing performance. There was also some evidence suggesting that more fixations on the forward part of the runway at certain times (40–30 range feet, 10–5 range feet) were linked with better landing position. All these results confirm that experts can make a quicker and more effective attentional adjustment during the landing process. This study offers a new way to understand the landing process and these conclusions will be of great value to develop the training course for new pilots to improve their landing skills.

Keywords: Landing · Experts and novices · Eye movement

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 308–320, 2019.
https://doi.org/10.1007/978-3-030-22507-0_24

1 Introduction

The landing process is critical for aviation safety. In the last ten years, 48% of fatal aviation accidents occurred in the approach and landing phases [1]. In recent research of Airbus, the top cause of hull lose in 1997–2017 was runway excursion and overshoot, which causes 35% of such accidents, followed by runway abnormal contact, with a percentage of ten [2]. Landing problems, including heavy landing, long-distance flare, deviation from the center line and landing with crab, was among the top causes (22%) of all abnormal aircraft status. Therefore, it is significant to study the influencing factors of landing performance for preventing flight accidents.

Human factor issues played an essential role in an unsafe landing. According to the report of [3], pilot manual handling errors accounted for 37% of errors in 2016 in commercial aviation accidents involving large aircraft in the world. Among these human errors, problems related to visual attention is a leading cause. A simple search on recent Chinese flight accident investigation reports can find that a large number of landing incidents were caused by problems related to pilots' visual attention, such as "Incorrect attention allocation when correcting direction deviation" at an accident of propeller and landing gear damage [4], "Captain's occupied attention to runway while flaring caused loss of aircraft states" at an incident of Airbus320 runway overshoot [5], "The late discovering of the over-speed descending rate and the inaccurate judgement of the height" at an incident of Boing737 tail strike [6], "Ignoring pitch control and misallocated attention were important causing factors in the incident" at an incident of Boing 737 dangerous hard landing [7].

Given the fact that pilots rely heavily on visual information and eye tracking is a useful tool to understand visual attention of pilots [8], some studies have found a significant correlation between the eye movement characteristics and landing performance during the landing phase [9–11]. However, only a few studies paid attention to the eye movement at different sub-stages of the landing. The current study sought to investigate how experts and novices differ in their eye movement characteristics during the landing process and how these characteristics are related to the landing performance.

1.1 Pilot' Cognitive Process During the Landing

The landing phase is a continuation of the final approach, in which the aircraft normally descend at a steady rate (usually 700 feet per minute) with a stable pitch attitude. When

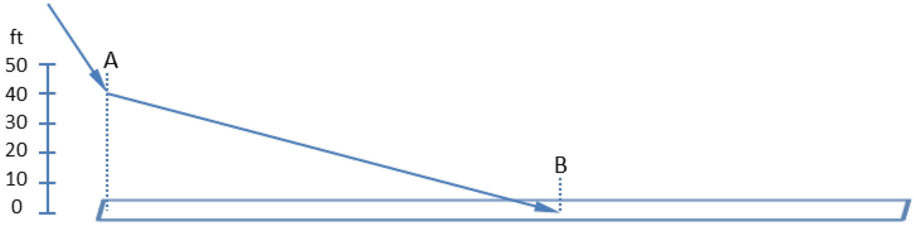


Fig. 1. The aircraft trajectory during the landing process

reaching the threshold area (i.e., around the A point in Fig. 1) at about 40 ft above ground for large aircraft, the pilots initiate the “flare” stage by increasing the pitch attitude gradually from 2 to 6°. The rate of descent also gradually decreases to less than 150 feet per minute during the flare until the plane touches the ground (the B point in Fig. 1). For large aircraft such as Airbus 320, the flare stage only takes about 8 s.

During this very short but critical period, the pilot needs to pay attention to a great deal of visual information which is both complex and fast changing. The success of landing depends heavily on whether the pilots can make a correct judgment about the vital information including altitude, attitude, speed, height and so on based on both flight instrument as well as external visual cues [11]. Since our attentional resources are very limited [12], how to allocate these finite resources is very important for a successful landing. Notably, before the flare, pilots should focus on the runway threshold area. This area, at the very beginning of the runway, containing many stripes and other clues as visual references, can help pilots judge the height and movement of the aircraft. Then the pilot needs to judge whether the trajectory of the aircraft is toward the aiming point (with two prominent white markings) using all available visual information. If a wrong judgment is made, they have to make additional effort to adjust the trajectory and attitude by operating the control system and the thrust system. Since the time margin is minimal during the flare stage, such a wrong judgment and corresponding readjustment may increase the likelihood of unsafe landing. As the flare continues, the pilot should change their viewpoint by moving their eyes “up” to see the rear part of the runway in order to predict the touchdown position. Viewpoint adjustment also helps to examine and adjust the yaw and bank angle by looking along the runway extension line. If such an adjustment of view is made insufficiently, the touch down will not go smoothly.

In the previous section, we have described what should be done during the landing process, in which the adjustment of viewpoint was a critical aspect. However, what is done by pilots is more important for both research and practical purposes. For achieving that goal, research on pilot’s eye movement can offer great help.

1.2 Eye Movement Tracking in the Cockpit

The research of pilots’ eye movement can be dated back to the 1940s, and the well-known concept of T type instrument layout was among its initial contributions [13]. With the development of computer technology in the late 1970s, a great deal of work has been done on pilots’ visual behavior [14–16]. Tracking the fixations of pilots can help understand what the pilots are interested in [17, 18]. This is because attentional switch can be reflected in the eye movement [19]. It has also been argued that covert attention can be detected through explicit ocular movements [20]. Researchers also found that people tend to look at objects that attract their interests; on the other hand, subjects could not pay attention to one object but look at another location [21]. More importantly, eye tracking can provide information that cannot be collected through operators’ self-report [22]. In a landing simulation experiment, researchers found that pilots of the air force were not able to report what they looked at during the process [23].

A recurrent issue in the research of the pilots' eye movement is to examine the difference between experienced pilots (experts) and novices. Expert pilots (those with more flight experience), similar to professionals in other areas, perform significantly better than novices, a natural improvement that comes with practice [24]. As they also have remarkable different eye movement patterns, a simple yet practical rationale behind this stream of research is that if we fully understand the eye movement patterns and corresponding cognitive states of the experts, we can find out a better way to train novices to think and act like them [25]. Here we would review the major findings on eye movement researches that focus on the landing stage (see different reviews made by [22, 26–28]).

Many studies found that experts and novices differ in the number of fixations (and dwells) and total duration. For example, Kasarskis and colleagues found that experts had significantly shorter dwells but more total fixations than novices. Another study also found that the performance of expert was better than novice; also, the experts had shorter fixation time, more fixation points, faster scan velocity, greater scan frequency and wider scan area than the novices. Sullivan et al. also found that more experienced pilots had much shorter dwells but more frequent view changes [29]. In a recent study, researchers found similar results that with less course deviation, more appropriate roll and pitch angle, the expert also showed shorter fixation time, smaller pupil size changes, lager scanning range, faster scan velocity, greater scan frequency and greater fixation frequency [30]. All these findings suggested that experts are more effective in information processing; thus they can acquire more information while the sampling cost on each piece of information is low.

A more useful and more important issue is to examine which areas are looked at by pilots during landing (the area of interest, AOI). A study found that experts had more fixations on the aiming point and airspeed but fewer fixations on the altimeter, as compared to the novices [29]. Another found that pilot paid more attention to the first half of the runway and tended to follow the threshold as it moved on the screen, as compared to other visual scene features. They also found that, for landing in the daytime, the supplementary out-of-cockpit visual cues could facilitate glide slope control performance [31]. Researchers found that in the process of landing, as compared to the low-performance group, the high-performance group had more fixations and longer fixation transfer times on the outside view, but fewer fixations and shorter fixation transfer times in the inside views [11]. Although the studies on the AOIs of landing are relatively limited, it suggests that the outside views (especially the runway) seem to be more important during the landing process.

While previous studies have made important discoveries in understanding the landing process, one crucial inadequacy is that they did not examine how the AOIs changes during the process (i.e., at different altitude). As described earlier, the areas that pilots need to see is itself subject to change during the landing process. Whereas at the end of the final approach and the beginning of flare, the front part of the runway of great importance, as the process continues, information on the rear part of the runway becomes increasingly vital. In this way, a natural adjustment of views seems important for a successful landing; however, such a pattern has not been observed in the previous eye tracking studies. As a result, the first purpose of the current study was to examine whether this adjustment process exists during the landing process. Moreover, since the experts are more adept in grasping the key information during this process, we expected that they could show a more salient adjustment as compared to the novices. Also, we also sought to see how eye movement characteristics at different sub-stages of landing are correlated to the performance criterion.

2 Method

2.1 Participants

Seventy-two pilots of civil airlines with valid commercial license participated in this study. Nobody had participated in similar experiments before. We excluded 15 participants since 11 had sampling rates below 75% and 4 had a severe line-of-sight deviation (i.e., judging from the position of the subjects fixation point, the line of sight is entirely out of the reasonable position to be watched), thus remaining 57 for the final analysis. Among them, the 29 expert pilots all had valid A320 class B instructor licenses. In average, they were 39.1 years old (SD = 4.10), with a mean flying time of 13, 921.4 h (SD = 3374.3). The average age of the 28 novices was 23.6 (SD = 1.50) and their average flight time was 267.2 h (SD = 21.79). All participants had good vision and did not need to wear any eyesight correction equipment. All participants took part in the study voluntarily.

2.2 Apparatus

The experiment was performed in the Airbus 320 full-motion flight simulator (CAE) Such a high fidelity simulator had its flight control, and display interfaces the same as that of the A320 aircraft, as shown in Fig. 2. The attitude, bank angle, crab angle and other data of the aircraft were taken from the flight simulator. The eye movement data were collected using the Tobii Glasses 2 eye-tracking system, as shown in Fig. 3. The sampling rate of the eye tracker was 50 Hz.

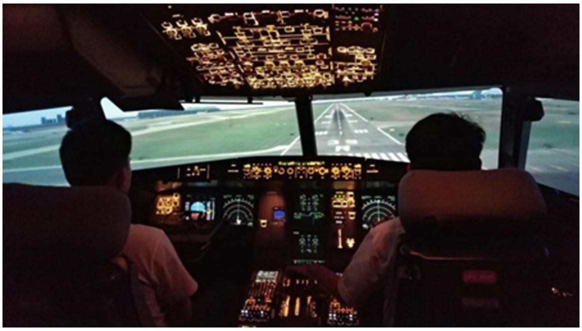


Fig. 2. Airbus 320 full-motion flight simulator

Fig. 3. Tobii Glasses 2 eye-tracker

2.3 Procedure and Task

Upon arrival, all participants signed the informed consent form and reported demographic information including their age and flying time. Afterward, the participants entered into the cockpit, put on the eye tracker and completed the calibration. After all participants were well seated, the experimenter initiated the task.

The aircraft was positioned at the take-off position at Pudong Airport (ZSPD), runway 35R. The environment conditions were: clear weather, 12:00 clock, temperature 15 °C, air pressure QNH 1013, wind calm, the brightness of runway approach light was set to level 3, and PAPI (approach gradient indicator) light wason. The aircraft was 64 tones, with a center of gravity 28%. In the cockpit, the light was set on the DIM (dark) position through the DOME (top light).

The participants were asked to perform a series of landing missions. They first needed to take-off, then turn left or right to join the visual pattern, then completed the visual approach and landing. The first mission was an exercise to familiarize them with the simulator operation. The data for the second and third missions were used for further analysis.

Data were processed by the eye-tracker Tobii I-VT Gaze Filter Sifting software. The runway area was divided into 2 AOIs according to the runway markings, the front part (from the beginning of the runway to the red line) and the rear part (from the red line to the end of the runway) as shown in Fig. 4.

2.4 Landing Performance

The landing performance was evaluated by six dimensions: landing position (latitudinal and longitudinal), handling stability, landing load, landing attitude, landing yaw angle and landing bank angle. Handling stability and landing load scores were obtained from an on-site rater during the experiment. The rest four dimensions were evaluated by two raters after the experiment by viewing videos recorded during the whole process with

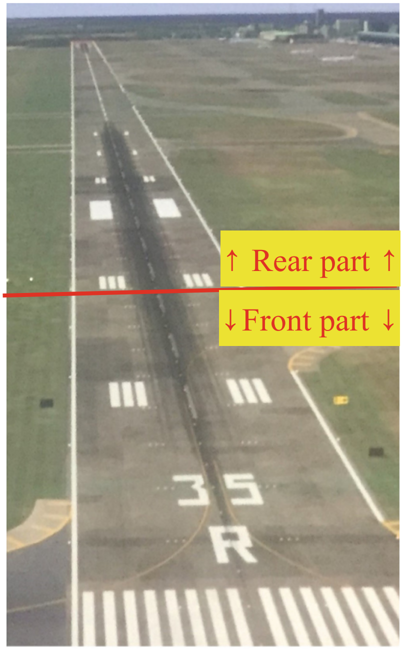


Fig. 4. Areas of interest on the runway

accurate feedback on the landing moment values. All dimensions were rated on a 5-point rating scale where one indicates very bad, and 5 indicates excellent. We used two measures to reflect the overall performance, the average score, and the lowest score. The latter was used because it was a daily routine in evaluating pilot performance during the training.

3 Results

3.1 Pilot Performance Between Experts and Novices

To test the difference of landing performance between experts and novices, we first summed all six sub-dimension scores and two overall measures scores over two flights and performed a series of independent sample t-tests on these dependent variables. We found that experts had significantly better scores in terms of landing position, landing load, handling stability, the average score, and the lowest score. However, the difference between experts and novices on attitude, yaw angle and bank angle of the aircraft were only approaching significant. Detailed information can be seen in Table 1.

Table 1. Comparison of performance dimensions between experts and novices

	Experts mean (S.E.)	Novices mean (S.E.)	t	p
Landing position	8.63(.15)	5.72(.31)	8.733	<.001
Landing load	8.51(.18)	6.44(.21)	7.624	<.001
Handling stability	8.29(.17)	5.56(.22)	9.973	<.001
Attitude	9.80(.08)	9.53(.13)	1.752	.084
Yaw angle	9.97(.03)	9.81(.08)	1.870	.066
Bank angle	10.00(.00)	9.91(.05)	1.874	.065
Average score	9.20(.07)	7.83(.11)	10.963	<.001
Lowest score	7.91(.16)	4.75(.24)	11.199	<.001

3.2 Fixation Adjustment and the Influence of Expertise

To examine how experts and novices adjusted their viewpoints during the landing process, we conducted a 6 * 2 * 2 repeated measures ANOVA using fixation amounts as the dependent variable. Height (50–40 ft, 40–30 ft, 30–20 ft, 20–10 ft, 10–5 ft and 5–0 ft) and fixation position (the front part vs. the rear part) were within-subjects factors, while expertise (expert vs. novice) was a between-subjects factor.

The results showed that the main effect of position was significant, F(1, 55) = 126.900, p < .001, the rear part was watched more as compared to the front part. The main effect of height was also significant, F(3.815, 209.822) = 7.651, p < .001. In total, more fixations appeared during late landing. But there was a significant interaction between position and height, F(3.594, 197.684) = 24.881, p < .001. Whereas the gaze in the forward part of the runway decreased as the landing continues, the fixations on the rear part of the runway increased. This pattern fully supports our adjustment hypothesis.

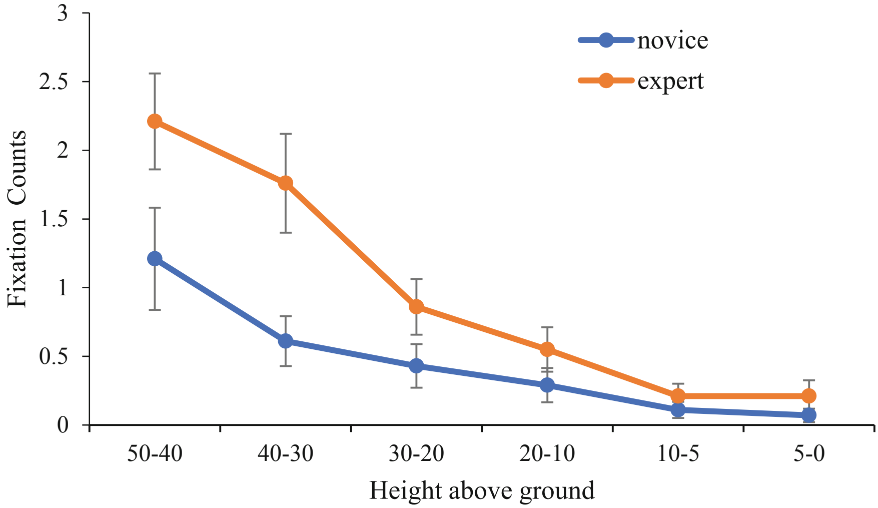


Fig. 5. Amount of fixations in the forward part of the runway

The main effect of expertise was not significant, F(1,55) = .551, p = .461. However, there was a significant three-way interaction among position, height and expertise, F(3.594,197.684) = 2.903, p = .028. We plotted this interaction in Figs. 5 and 6. While both experts and novices showed the adjustment pattern of fixations (as getting closer to the ground, fixations in the front part reduces while that in the rear part increases), the adjustment of experts was more salient. This pattern can be seen by the fact that the expert had a steeper drop line in Fig. 5 and a steeper ascending line in Fig. 6. Thus it supports our hypothesis that the experts are quicker and better at making this adjustment.

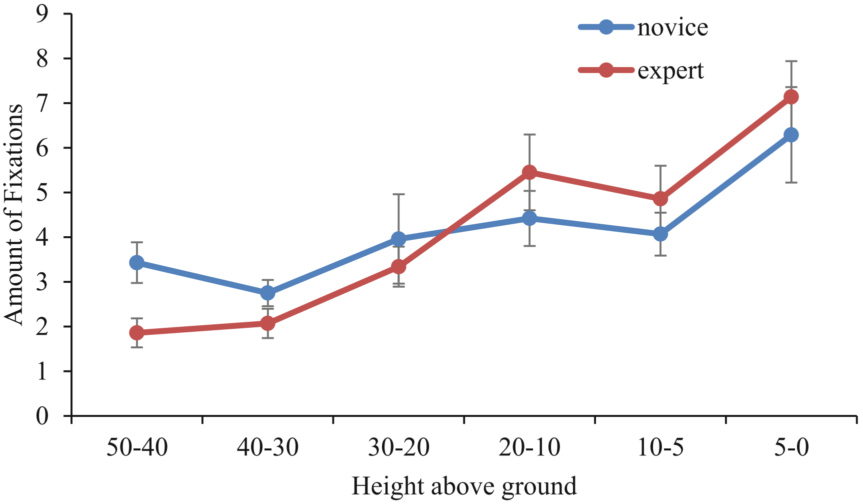


Fig. 6. Fixations in the rear part of the runway

3.3 The Correlations Between Pilot Performance and Fixations

To further explore the relationship between pilot performance and fixation counts, we conducted correlational analyses among all performance indicators and fixation counts. Results showed early fixations on the rear part (higher than 20 ft) of the run way was linked with worse scores on multiple performance indicators. There was also some evidence suggesting that more fixations on the forward part of the runway at certain times (40–30 ft, 10–5 ft) were linked with better landing position. Details were presented in Table 2.

Table 2. Thecorrelations between fixation amount and pilot performance scores

Performance criterion	Eye fixation times at the front part						Eye fixation times at the rear part					
	50–40 ft	40–30 ft	30–20 ft	20–10 ft	10–5 ft	5–0 ft	50–40 ft	40–30 ft	30–20 ft	20–10 ft	10–5 ft	5–0 ft
1. Landing position	.233	**.307***	.138	.193	**.277***	.189	**−.276***	**−.371****	**−.320***	−.153	−.055	−.069
2. Landing load	.098	.148	.205	.159	**.274***	.151	**−.382****	−.240	−.182	−.066	−.218	−.031
3. Manipulation stability	.130	.237	.247	.231	.154	.090	**−.470****	−.210	−.138	.038	−.060	.068
4. Attitude	.190	.115	.071	−.040	.020	−.003	−.080	−.134	−.153	−.235	.000	−.026
5. Yaw angle	.072	.010	−.211	−.184	.125	.096	.072	**−.282***	**−.594****	**−.264***	−.121	**−.376****
6. Bank angle	−.195	.077	−.004	.129	.091	.070	**−.324***	−.084	.018	−.083	.034	**−.316***
7. Average score	.184	**.261***	.190	.187	.255	.157	**−.396****	**−.327***	**−.286***	−.119	−.112	−.057
8. Lowest score	.160	.252	.151	.159	.229	.186	**−.379****	**−.286***	**−.299***	−.099	−.111	−.007

Note: 50–0 ft means the fixation times were counted during the time during which the aircraft was descending from 50 ft high to 40 ft high, and so on. *indicates $p < .05$, **indicates $p < .01$.

4 Discussion

In the study, we had three purposes: first, we wanted to examine whether there was a fixation adjustment during the landing process; second, whether experts and novices differ in their adjustment patterns; third, we also wanted to identify certain fixation indicators that might be linked with landing performance. By using a large sample of professional pilots as participants, we investigated these questions on a high fidelity flight simulator.

We first found evidence that there exists a fixation adjustment during the landing process. More specifically, the pilots reduce their times of watch to the front part of the runway but increase their times of watch to the rear part of the runway during the landing process. Although this pattern is entirely in align with operation manuals, to the best of our knowledge, this paper is the first to make that report. The reason is that we are among the very few to make a real process analysis by observing the behavior patterns at different sub-stages of the landing.

Based on this, we went further to examine the effect of expertise and found more interesting discoveries: although all pilots adjust their fixations during the landing process, the adjustment made by experts was more salient, as compared to the novices. Experts put much more attention to the front part of the runway from the very beginning to almost halfway (50–20 ft). During this critical stage, looking at the entrance of the runway can help find the proper start point of flare. Since early flare may lead to long landing and loss of runway while late flare may lead to a hard landing, a precise evaluation of the start point is valued by the experts and put more focuses on it. However, later their fixations quickly dropped to be no different as compared to the novices. On the contrary, experts paid significantly less amount of attention to the rear part of the runway at the beginning, but their focus to the rear part increases steadily until it takeover the novices'. All these results confirm our proposal that experts can make a quicker and more effective attentional adjustment during the landing process.

Besides, we also find there was some relationship between the fixations at each sub-stages and the final landing performance. A quite clear finding was that earlier occupation with the rear part of the runway might be harmful to landing safety. The risk of this watch pattern might be because, at the early stages of landing (higher than 30 ft), the limited attentional resources must be devoted to more important areas at the front part of the runway, the shift of attention to the rear part, either by self-directed or cue-elicited reasons, can be considered as a distraction. Future studies may go further to investigate what kind of method can be useful in reducing such unwanted gazes.

Several limitations must be mentioned before making the conclusion. Although we used a large sample of pilots and their data were collected from a high fidelity simulator, we only examined their behavior in a basic and simple setting. For generalize the findings and gather more information, future studies may benefit from investigating other environmental conditions such a wind, darkness, fog, etc. Second, the landing performance was evaluated by subjective evaluations rather than by objective measures. However, although there are some objective measures of landing performance such as weighted latitudinal and longitudinal positions [24], and landing distance, load and pitch angle [11], these indicators may not fully grasp the essence of landing safety which is sensed more accurately by human operators. Another problem is that any failure in a single dimension can lead to disastrous consequences, so an overall performance indicator with an emphasis on the worst performance dimension is required. As a result, this study used subjective evaluation as the performance measure.

5 Conclusion

The current study found evidence suggesting the pilots do adjust their gaze points during the landing process and the experts made such an adjustment in a more salient manner. Moreover, earlier fixations on the rear part of the runway may result in bad landing performance. This study offers a new way to understand the landing process and the findings are of great value to develop new training courses for new pilots to improve their landing skills.

Acknowledgments. This research was supported by National Key Research and Development Plan [grant number: 2016YFB1001203] and the Natural Scientific Foundation of China [31671148]. We thank Ms. Xiaolin Jiang and Mr. Jian Gong and other technicians of TOBII who offered so much help in data collection. We are also grateful to the more than 70 captains and co-pilots of CEA for their active participation in this experiment, Thanks to Mr. Ji Liu, Mr. Jian Gu, Mr. Yi Zhou, Mr. Shengrui Zhou, and so on, for their great assistance during the experiment.

References

1. Boeing: Statistical Summary of Commercial Jet Airplane Accidents Worldwide Operations-1959–2016, Aviation Safety, Boeing Commercial Airplanes (2017)
2. Airbus: A Statistical Analysis of Commercial Aviation Accidents 1958–2017, Safety First, 22 June 2018
3. IATA: Safety Report 2016, pp. 139–141, 53rd edn. The International Air Transport Association (2017)

4. CAAC: Investigation report on the accident of propeller and landing gear damage caused by a bounce during the landing of aircraft DA40D/B-9515 at Chaoyang Flight Academy on 25 March. CAAC Liaoning Provincial Regulatory Bureau (2014)
5. CAAC: Investigation report on runway incident at Jingdezhen Airport by Shenzhen A320/B-6312. CAAC East China Regional Administration (2014)
6. CAAC: Report on the investigation report on the serious accident of the tail strike of flight MF8381 of Xiamen Airlines at Baiyun Airport on April 12, 2016. CAAC Central South Regional Administration (2016)
7. CSH: Investigation report on the serious hard landing of Flight FM857. Safety Supervision Department, Shanghai Airlines Co., Ltd. (2017)
8. EASA: Collision Avoidance. Safety analysis and research department, European Aviation Safety Agency (2010)
9. Liu, Z., Yuan, X., Fan, Y., Liu, W., Kang, W.: Comparison of eye movement behavior between experts and new hand during flight simulations of aircraft landings. Space Med. Med. Eng. **22**(5), 358–361 (2009)
10. Wang, L., Li, H., Dong, D., Shu, X.: Relationship between the technical skills and eye-movement indicators of pilots. In: Proceedings 19th Triennial Congress of the IEA, vol. 9, p. 14 (2015)
11. Wang, L., Ren, Y.: Study on fixation pattern in visual landing. Psychol. Sci. **40**(2), 283–289 (2017)
12. Wickens, C.: Engineering Psychology and Human Performance, 4th edn, pp. 3–4 (2013)
13. Fitts, P.M., Jones, R.E., Milton, J.L.: Eye movements of aircraft pilots during instrument-landing approaches. Aeronaut. Eng. Rev. **9**(2), 24–29 (1950)
14. Wu, X., Wanyan, X., Zhuang, D.: Pilot's visual attention allocation modeling under fatigue. Technol. Health Care **23**(s2), S373–S381 (2015)
15. Carroll, M., et al.: Enhancing HMD-based F-35 training through integration of eye tracking and electroencephalography technology. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2013. LNCS (LNAI), vol. 8027, pp. 21–30. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39454-6_3
16. Ferrari, F., Spillmann, K.P., Knecht, C.P., Bektas, K., Muehlethaler, C.M.: Improved pilot training using head and eye tracking system. In: Proceedings of the 3rd International Workshop on Eye Tracking for Spatial Research. ETH Zurich (2018)
17. Shinar, D.: Looks are (almost) everything: where drivers look to get information. Hum. Factors **50**(3), 380–384 (2008)
18. van de Merwe, K., van Dijk, H., Zon, R.: Eye movements as an indicator of situation awareness in a flight simulator experiment. Int. J. Aviat. Psychol. **22**(1), 78–95 (2012)
19. Klein, R.: Does oculomotor readiness mediate cognitive control of visual attention? Attent. Perform. **8**, 259–276 (1980)
20. Rizzolatti, G., Riggio, L., Dascola, I., Umiltá, C.: Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. Neuropsychologia **25**(1), 31–40 (1987)
21. Hoffman, J.E., Subramaniam, B.: The role of visual attention in saccadic eye movements. Percept. Psychophys. **57**(6), 787–795 (1995)
22. Glaholt, M.G.: Eye tracking in the cockpit: a review of the relationships between eye movements and the aviator's cognitive state. Scientific Report, DRDC-RDDC-2014-R153 (2014)
23. Robinski, M., Stein, M.: Tracking visual scanning techniques in training simulation for helicopter landing. J. Eye Mov. Res. **6**(2), 1–17 (2013)

24. Kasarskis, P., Stehwien, J., Hickox, J., Aretz, A., Wickens, C.: Comparison of expert and novice scan behaviors during VFR flight. In: Proceedings of the 11th International Symposium on Aviation Psychology, vol. 6 (2001)
25. Mourant, R.R., Rockwell, T.H.: Strategies of visual search by novice and experienced drivers. Hum. Factors **14**(4), 325–335 (1972)
26. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychol. Bull. **124**(3), 372 (1998)
27. Jacob, R.J., Karn, K.S.: Eye tracking in human-computer interaction and usability research: ready to deliver the promises. In: The Mind's Eye, pp. 573–605 (2003)
28. Ziv, G.: Gaze behavior and visual attention: a review of eye tracking studies in aviation. Int. J. Aviat. Psychol. **26**(3–4), 75–104 (2016)
29. Sullivan, J., Yang, J.H., Day, M., Kennedy, Q.: Training simulation for helicopter navigation by characterizing visual scan patterns. Aviat. Space Environ. Med. **82**, 871–878 (2011)
30. Xiong, W., Wang, Y., Zhou, Q., Liu, Z., Zhang, X.: The research of eye movement behavior of expert and novice in flight simulation of landing. In: Harris, D. (ed.) EPCE 2016. LNCS (LNAI), vol. 9736, pp. 485–493. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_47
31. Kim, J., Palmisano, S.A., Ash, A., Allison, R.S.: Pilot gaze and glideslope control. ACM Trans. Appl. Percept. (TAP) **7**(3), 18 (2010)

From Paper-Based Operational Procedures to Onboard Context-Sensitive Information System (OCSIS) for Commercial Aircrafts

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 manuals, Electronic Flight Bag (EFB) and electronic systems onboard to help them perform procedures to ensure safety, efficiency and comfort on commercial aircrafts. Although the manuals and procedures in EFB are not context-sensitive, management of interconnections among these operational documents can be a challenge for pilots, especially when time pressure is high in normal, abnormal, and emergency situations. This paper presents a possible solution for an on board context-sensitive information system (OCSIS) that would be an alternative to current electronic library systems and traditional paper-based onboard documentation systems. A concurrent analysis of existing on board documentation content, was carried out to determine what are the main requirements for such OCSIS. An analysis of context in commercial aviation is proposed and applied to a possible OCSIS solution. This research and design work presents a methodology that supports human-centered design (HCD) of onboard context-sensitive information systems. We developed and tested the OCSIS in the context of commercial aircraft flight decks. Although several findings were elicited from the various testing sessions that we had during the design cycle of this academic work, the main contribution is the articulation of various techniques and tools that make HCD feasible and effective.

Keywords: Commercial · Aircraft · Onboard information system · Human-centered design · Workload · Context

1 Introduction

Civil aviation is a rich industry, which is based on constant innovations on technology, organization and practices to improve safety, efficiency, and comfort [1]. Consequently, ICAO and national regulatory organizations have to incrementally develop appropriate standards. Flight crews follow Standard Operating Procedure (SOP) to complete tasks and ensure safety. In abnormal situations such as a malfunction of an aircraft system or extreme weather condition, abnormal procedures have to be used for trouble-shooting.

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 321–331, 2019.
https://doi.org/10.1007/978-3-030-22507-0_25

Operational documentation is regularly improved using experience feedback for normal, abnormal, and emergency situations [2]. 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 [3].

These documents have been paper-based since the beginning of aviation. Today, the concept of an Electronic Flight Bag (EFB) has been developed. "An Electronic Flight Bag is simple and attractive: a pilot's personal flight-deck computer. Airlines are eager to have customized EFBs on board, and manufacturers are eager to develop and supply them. Software providers are eager to customize software for EFBs as well. There are multiple concepts for what an EFB is, ranging from low-end to high-end devices" [4].

For the last two decades, the shift from paper to electronic documentation has been discussed, modeled, and partly operationalized [5]. This shift is not a matter of transferring paper-based information onto an electronic format (e.g., PDF format). Electronic support offers different kinds of capabilities that paper cannot offer (e.g., context-sensitivity). In other words, onboard paper-based documentation can be connected to flight parameters to provide useful static knowledge, which pilots need to contextualize during operations. Electronic support provides capabilities that enable connectivity, and consequently the provision of dynamic information in context. We claim that electronic documentation, renamed "onboard information system (OIS)," contributes to improving the perception and comprehension of the current situation, as well as supporting decision-making and action-taking.

2 Re-organized Procedures with Respect to Context

The aeronautical community is constantly working on transferring related information from paper to computer support (e.g., Airbus's Onboard Information System and Boeing's Electronic Flight Bags). These new tools cover aircraft technical information, operating manuals, performance calculations, and mission management information. They are context-free databases. However, computer support enables interconnectivity among relevant pieces of information (i.e., hypertext links) and between cockpit information and flight parameters (i.e., context-sensitivity) [6].

This dissertation presents a new system, the Onboard Context-Sensitive Information System (OCSIS), which is available on a tablet wirelessly connected to relevant cockpit parameters. OCSIS enables and requires new information formatting (e.g., the concept of page is no longer relevant). In addition, OCSIS's internal information is structured with respect to context.

"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" [7]. Context-aware computing was first discussed by Schilit and Theimer in 1994 as software that "adapts according to its location of use, the collection of nearby people and objects, as well as changes to those objects over time" [8]. Since then, Dey defined "A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task" [7, 9].

Context patterns can be classified into three categories: independent, hierarchically dependent, and interdependent (Fig. 1) [10]. By a conjunction of situational conditions, the context pattern equals to situational-condition-1, plus situational-condition-2, until plus situational-condition-X. In the sub-context, the relationship can be hierarchically dependent or interdependent.

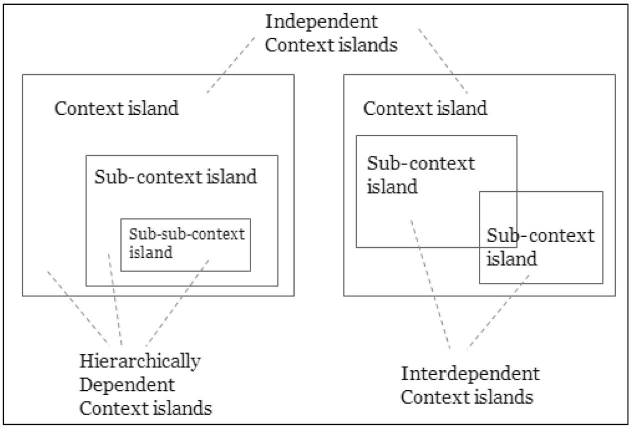


Fig. 1. Relationship of context islands [10]

3 Human-Centered Design Approach

Human-Centered Design (HCD) has been used to incrementally improve OCSIS toward an acceptable mature version (i.e., incremental prototype development, test, and modification) [11]. 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 (see Fig. 2) [12]. 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.

The design of a system is never finished even though at some point, delivery is required. This is why maturity has to be assessed [13]. The more OCSIS is being used and tested, the newer cognitive functions emerge and need to be taken into account either in system redesign, system training, or operations support [12].

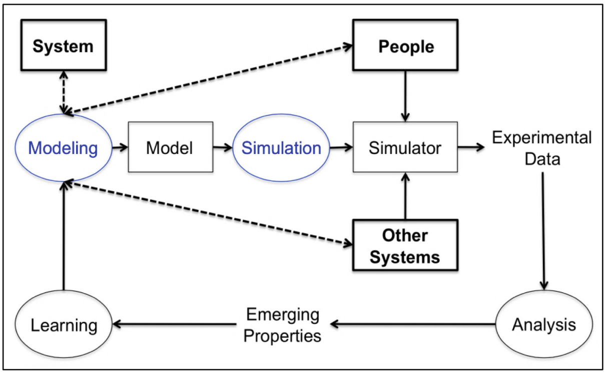


Fig. 2. Human-centered design approach [12]

During the early stage of design, it is very important to have professional pilots involved to set up a high-level prototype correctly. They need to describe stories applicable to the product being developed and how these historical events may be induced during the product operational phase. This is an important role of participatory design. In the case of OCSIS, professional pilots participated in the design process from the beginning and in all testing sessions (i.e., formative evaluations), providing experience feedback toward improving design.

4 Scenario-Based Design

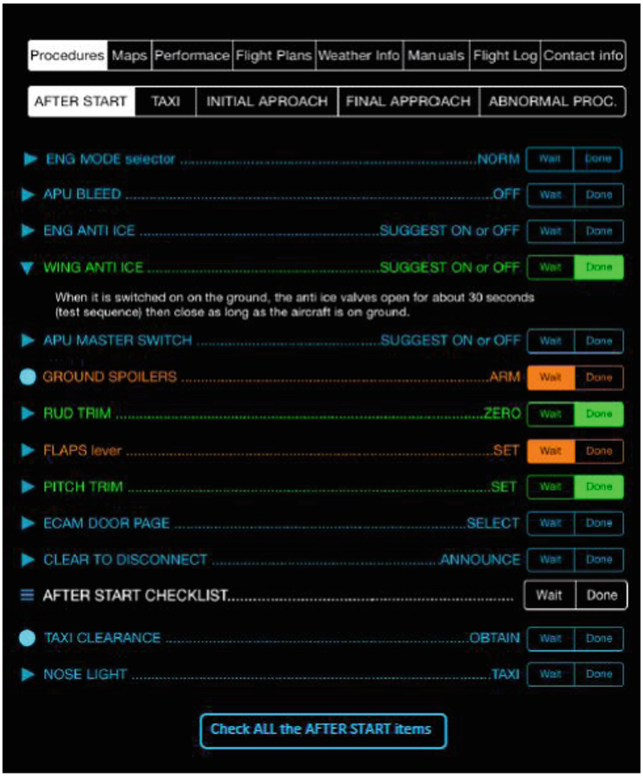
Scenario-based design requires descriptions of how people use technology, and discussing and analyzing how the technology is used to reshape their activities [14]. It is carried out during the early stages of the design process. During the OCSIS design process, a prototype was incrementally developed and tested by means of using usability principles and criteria.

Scenarios can be abstracted and categorized, helping designers to recognize, capture, and reuse generalizations, and to address the challenges that technical knowledge often lags behind when considering the needs of technical design [15].

Since airlines are allowed to equip tablets on the flight deck to replace heavy manuals and charts, we have chosen to implement OCSIS on a tablet. Another requirement for OCSIS is pilots' need for assistance in problem-solving under high time-pressure work. In consideration of this important aspect, we developed scenarios that enabled us to test the value of OCSIS in such situations. OCSIS scenarios were developed keeping in mind normal and abnormal situations, as well as training and instructing. We then developed scenarios for the following contexts: (1) Fuel Leak, (2) Descent, (3) Approach, (4) Flaps Locked, and (5) Landing. These scenarios were used in Human-in-the-loop simulations (HITLS) with professional pilots for OCSIS tests. This was the first step of OCSIS's scenario-based design.

5 Design of OCSIS Prototype

OCSIS is currently programed 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. The default/initial page displays procedures and actions that crews need to perform or have performed (see Fig. 3). Parts of normal scenarios are chosen to apply in OCSIS, such as After Start procedures. "Ready to do" actions are in cyan. Once the action is completed and OCSIS can access the related parameters' status, it automatically becomes "green." The title of flight phase, normal checklists and more information for actions are in white. Postponed actions or checks and the title of abnormal procedures are in amber [16].



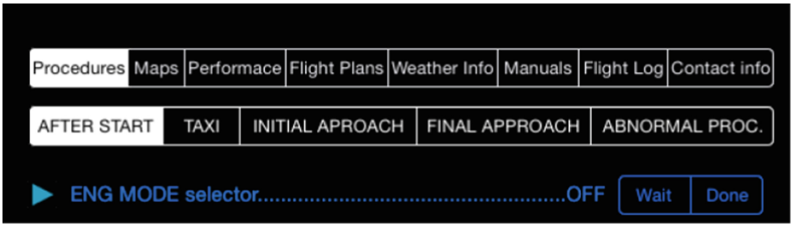
ENG MODE selector ....NORM
APU BLEED ....OFF
ENG ANTI ICE ....SUGGEST ON or OFF
WING ANTI ICE ....SUGGEST ON or OFF

When it is switched on on the ground, the anti ice valves open for about 30 seconds (test sequence) then close as long as the aircraft is on ground.

APU MASTER SWITCH ....SUGGEST ON or OFF
GROUND SPOILERS ....ARM
RUD TRIM ....ZERO
FLAPS lever ....SET
PITCH TRIM ....SET
ECAM DOOR PAGE ....SELECT
CLEAR TO DISCONNECT ....ANNOUNCE
AFTER START CHECKLIST........
TAXI CLEARANCE ....OBTAIN
NOSE LIGHT ....TAXI

Fig. 3. OCSIS's normal procedures interface (Color figure online)

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 (see Fig. 4).

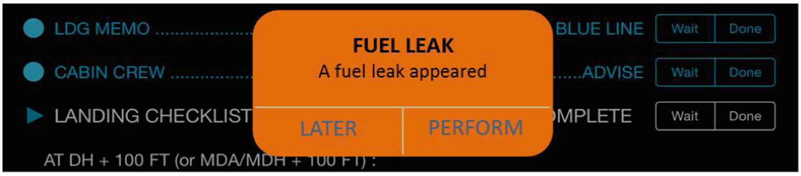


ENG MODE selector...OFF

Fig. 4. OCSIS's menu

- Procedures: Normal procedures are categorized by flight phases, and a flight phase icon enables manual selection of the current flight phase. Once selected, this icon becomes green automatically. Otherwise, pilots are required to check it into green manually. In the case of an abnormal situation, pilots can select abnormal procedures or stay in normal procedures.
- Maps: A list of Jeppesen Charts is available. Their sequence can change in real-time with respect to aircraft location.
- Performance Charts: Four categories of performance charts are available in pdf format: Takeoff Analysis PERF, Inflight PERF, Quick Reference PERF, and FCOM. Pilots can choose charts manually.
- Flight Plan: The current flight plan is in pdf format for pilots to check and review.
- Weather Info: Real-time weather is provided to pilots through Internet connection.
- Manuals: all manuals can be listed in pdf formal. Pilots can refer to any manuals they need.
- Flight Blog: Pilots can type the problem or observation unusually during the flight for ground maintenance to review and check, include flight date, flight number, departure, arrival, and pilots' names.
- Contact Info: The nearby Air Traffic Controller (ATC) contact frequencies are displayed for pilots to get them fast.

The current version of OCSIS includes two abnormal scenarios, "Fuel Leak" and "Flaps Locked". Context patterns trigger procedures in real-time both in normal and abnormal situations. In an abnormal situation such as "Fuel Leak," OCSIS will immediately inform the pilot about this malfunction by displaying a pop-up information window (see Fig. 5). Pilots can become aware of the problem through the pop-up window and start following actions, which directs to additional "Fuel Leak" procedures (see Fig. 6).



LDG MEMO ....

CABIN CREW ....

LANDING CHECKLIST

AT DH + 100 FT (or MDA/MDH + 100 FT) :

BLUE LINE

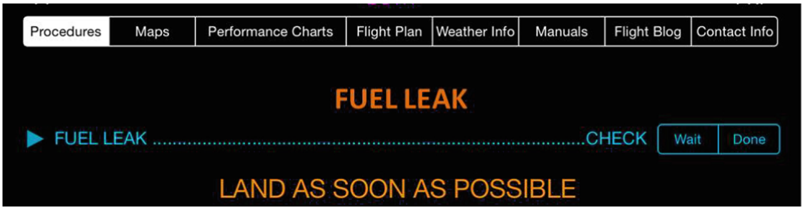
.....ADVISE

MPLETE

FUEL LEAK

A fuel leak appeared

Fig. 5. "Fuel Leak" triggering



FUEL LEAK

▶ FUEL LEAK ....CHECK

LAND AS SOON AS POSSIBLE

Fig. 6. "Fuel Leak" page

6 Pilots' Workload Evaluation

The first series of tests of Onboard Context Sensitive Information System (OCSIS) were carried out in three steps with respect to three main topics: Pilot-OCSIS interaction, situation awareness properties of OCSIS (i.e., OCSIS's capacity to provide the right information at the right time with the right format – this provides pilots with situated perception and understanding of what is going on and what should be done accordingly), and OCSIS's location [16]. Based on HITLS from the beginning of the design process, OCSIS was incrementally improved after each test based on the professional pilots' feedback. With the improved OCSIS set in the simulator (see Fig. 7), pilots' workload tests were taken in HCDi lab. We invited five pilots to take the experiment. They used paper-based manuals and OCSIS randomly during the experiment in two abnormal scenarios, Fuel Leak and then Flap Locked. Some pilots used manuals for Fuel Leak and OCSIS for Flap Locked, and some pilots used OCSIS for Fuel Leak and manuals for Flap Locked.

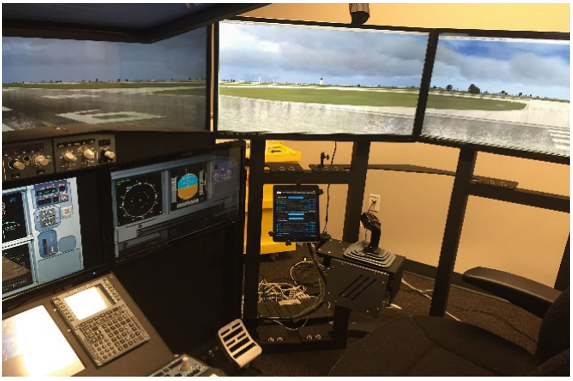


Fig. 7. The set of iPad's location

NASA-TLX criteria are used to measure workload in the two sessions to compare [17, 18]. Pilots are required to grade their workload and performance by values from 1 to 10. Based on the grading values collected from four pilots, the overall workload of dealing with each failure is calculated (see Fig. 8).

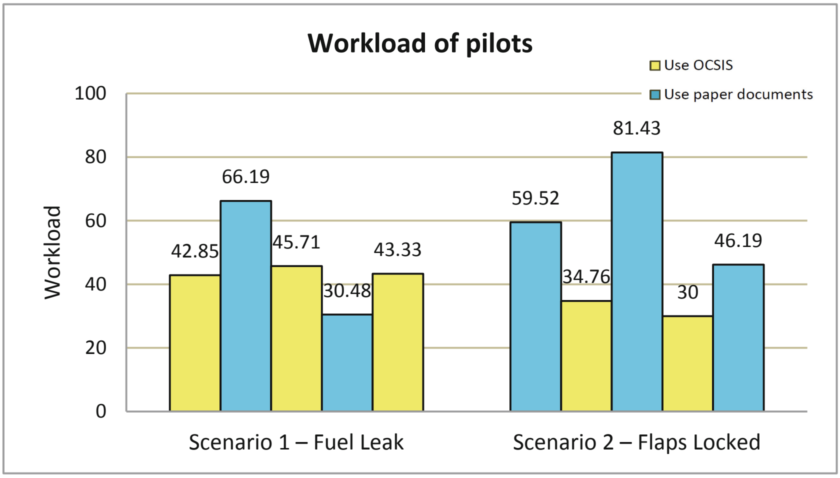


Fig. 8. Overall workload of pilots (Color figure online)

The blue cylinders represent workload by using paper-based documents, and the yellow cylinders represent workload by using the OCSIS under the context of Fuel Leak and Flaps Locked. Analyze the sample mean and standard deviation of pilots' workload by using two different references (see Table 1), we can find that these five pilots' workload is less when they used OCSIS in the two abnormal scenarios. As an assistance for pilots during performing procedures, OCSIS can provide them standard information immediately which can save them a lot of time on searching procedures and calculating performance data. However, it still needs plenty of samples to verify if OCSIS can reduce pilots' workload in more scenarios.

Table 1. Workload analyses

	Numbers	Sample mean	Standard deviations
Workload-OCSIS	5	39.33	5.950271
Workload-Manuals	5	56.762	17.36384

7 Discussion

At this point, it is important to mention that this dissertation is not a classical human factors and ergonomics research project where human issues are discovered after engineering work is done (i.e., corrective ergonomics). It is, instead, a human-systems integration project that is based on expert analysis of previous experience provided by accident reports and professional pilot expertise, creative design, iterative usability, usefulness assessments of prototypes using human-in-the-loop simulations involving professional pilots (i.e., formative evaluations), and constant creative (re)design of solutions. These solutions are not only technological, but also organizational and individual (in the sense of job roles and functions). The full capacity and ability of OCSIS will not be compared to current paper-based operational documentation, but this research is going to provide a first contribution to the maturity process of OCSIS design. Formative evaluation helps to correct design flaws form new design decisions. It uses heuristics and is cooperative (i.e., involves real users). This approach is strongly based on a meticulous follow up of design history, which involves purposes (such as the ones deduced from accident analyses, as shown above for example), solutions proposed (including various possible options), and criteria that make design choices possible and effective (including pilots' activity analysis in human-in-the-loop simulations).

We make a distinction between task (i.e., what is prescribed to be performed) and activity (i.e., what is effectively performed) [10, 13, 19]. Activity analysis is based on direct observation, questionnaires, interviews, and human factors studies. Today, activity analysis is possible because we have realistic modeling and simulation capabilities. It enables us to better understand how human operators execute tasks, and eventually defines emergent activities that are necessary to accomplish their goals. Performing activity analysis during design is one of the most important assets of human-centered design.

8 Conclusion

One of the most critical features of OCSIS is context-sensitivity, which enhances operational procedures following. Context-sensitivity provides pilots with more flexibility in flight operations. The main issue of this approach is the definition of context patterns. Indeed, getting the right procedure at the right time means that the system has the appropriate context pattern that matches the current situation and triggers the appropriate procedure. Consequently, context pattern correctness is crucial. In other words, context patterns have to recognize the right context and propose the right procedure to the pilot [11]. Future work will be dedicated to context pattern acquisition and formalization in the context of OCSIS development.

Creativity and evaluation are two important processes in Human-Centered Design (HCD). Creativity can be seen as integration of existing things [20]. Regarding OCSIS, we integrated well-known technology and techniques to create a new tool [11]. HCD recommends testing activity from the very beginning of design using realistic technology (e.g., realistic operating procedures and aircraft simulator), organization (e.g., two crewmen cockpit organization) and people (e.g., professional pilots). This approach led to very credible simulations, which were used to increase the tangibility of OCSIS (i.e., flexibility, maturity, and stability).

The shift from paper to electronics enables the emergence of context-sensitive HCI. Consequently, representation of context, identification of relevant contextual information, and actual interaction at the right time with the right information are key issues that needed to be further explored and developed. Our approach was based on creativity. We explored various kinds of solutions that attempted to improve safety, efficiency, and comfort on the flight deck. In future work, more scenarios need to be designed on OCSIS and more pilots are needed to take the experiments to improve the system.

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. Sudarsan, H.V.: Safety management principles. In: Workshop on the Development of National Performance Framework for Air Navigation Systems, Nadi, Fiji, 28 March–April 2011 (2011)
2. Ramu, J.-P., Barnard, Y., Payeur, F., Larroque, P.: Contextualized operational documentation in aviation. In: de Waard, D., Brookhuis, K.A., Weikert, C.M. (eds.) Human Factors in Design, pp. 1–12. Shaker Publishing, Maastricht (2004)
3. Tan, W.: From commercial aircraft operational procedures to an onboard context-sensitive information system. In: Proceedings of the HCI-Aero Conference, Santa Clara, CA. ACM Digital Library (2014)
4. Chandra, D.C., Mangold, S.J.: Human factors considerations for the design and evaluation of electronic flight bags. In: DASC 2000, 19th Proceedings of the Digital Avionics Systems Conference (Cat. No. 00CH37126), vol. 2, pp. 5A1/1–5A1/7 (2000). ISBN 9780780363953
5. Chandra, D.C.: Human factors evaluation of electronic flight bags. In: Chatty, S., Hansman, J., Boy, G. (eds.) Proceedings of the International Conference on Human-Computer Interaction in Aeronautics, pp. 69–73. AAAI Press, Menlo Park (2002)
6. Boy, G.A.: Indexing hypertext documents in context. In: Proceedings of Hypertext 1991, pp. 51–61. ACM Press (1991)
7. Dey, A.K.: Understanding and using context. Pers. Ubiquitous Comput. **5**(1), 4–7 (2001)
8. Schilit, B., Theimer, M.: Disseminating active map information to mobile hosts. IEEE Netw. **8**, 22–32 (1994)
9. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.: Towards a better understanding of context and context-awareness. In: Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing, HUC 1999, pp. 304–307 (1999)
10. Boy, G.A.: Cognitive Function Analysis, vol. 2. Greenwood Publishing Group, Westport (1998)
11. 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, pp. 9–14 (2015)
12. Boy, G.A.: From automation to tangible interactive objects. Annu. Rev. Control **38**, 1–11 (2014). https://doi.org/10.1016/j.arcontrol.2014.03.001. ISSN 1367-5788
13. Boy, G.A.: Design for safety: a cognitive engineering approach. In: Barnard, Y., Risser, R., Krems, J. (eds.) The Safety of Intelligent Driver Support Systems. Design, Evaluation and Social Perspectives. Ashgate, Farnham (2011). ISBN 978-0-7546-7776-5

14. Rosson, M.B., Carroll, J.M.: Scenario based design. In: Human-Computer Interaction, Boca Raton, FL, pp. 145–162 (2009)
15. Boy, G.A.: The Handbook of Human-Machine Interaction: A Human-Centered Design Approach. Ashgate, Farnham (2011)
16. Tan, W., Boy, G.A.: Tablet-based information system for commercial air-craft: onboard context-sensitive information system (OCSIS). In: Proceedings of the HCI Conference (2018)
17. Hart, S., Staveland, L.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock, P., Meshkati, N. (eds.) Human Mental Workload, pp. 139–183. North Holland, Amsterdam (1988)
18. Stanton, N.A., Salmon, P., Rafferty, L.A., Walker, G.H., Baber, C., Jenkins, D.P.: Human Factors Methods: A Practical Guide for Engineering and Design. Ashgate Publishing Co., England (2005)
19. Boy, G.A.: Orchestrating Human-Centered Design. Springer, London (2013). https://doi.org/10.1007/978-1-4471-4339-0
20. Boy, G.A.: Tangible Interactive Systems - Grasping the World with Computers. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30270-6

Airworthiness Compliance Criteria in Ergonomic Design of Cursor Control Device for Civil Aircraft

Lei Wu(✉) and Jian Xu

Shang Aircraft Design Research Institute, Shanghai 201210, China
wuleil@comac.cc

Abstract. Cursor control device (CCD) has been applied to the integrated avionics system in the cockpit of modern civil aircraft, and gradually becomes one of the main controls in the cockpit. From the perspective of operation frequency and workload, ergonomic design of cursor control device, especially the design of installation location, will become an important aspect of ergonomic airworthiness. Based on anthropometric data and ergonomics combined with biomechanical theory, the control posture requirement of CCD is determined by analyzing the operational situation of CCD, and the reasonable installation location of CCD is calculated by computer aided design technology. From the results of computational analysis and evaluation, there is theoretically a limited installation area, which can meet the needs of most pilots to operate CCD conveniently and efficiently over all flight phases. While satisfying the airworthiness provisions, the relevant theoretical methods and quantitative analysis results can be used as criteria to show the compliance to the authorities.

Keywords: Civil aircraft · Avionics · Cursor control device · Ergonomics

1 Background

With the rapid development of electronics and computer technology, the integrated degree of avionics system is getting higher and higher, and its functions are becoming more and more complex. Especially in interactive control, if each function is assigned an independent control, they will not only challenge the already constraint cockpit space, but also make pilots more difficult to operate.

Through the combination of virtual control technology and cursor control device, a new generation of integrated avionics system has the technical basis of providing pilots with more efficient, safe and convenient means of interaction. However, due to the late entry of new technology into the cockpit, further research on the ergonomic design of CCD, especially in the installation location, shape and size, and its quantitative airworthiness compliance standard is still needed. Traditional cockpit workspace design methods tend to use human body model data as the basis, and computer aided design tools (such as JACK) are used to analyze ergonomically the vision of existing design solution and the accessibility of control devices [1]. The analysis results can be used as the basis of iterative design or evidence of airworthiness. However, due to the limited

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 332–341, 2019.
https://doi.org/10.1007/978-3-030-22507-0_26

cockpit space and various devices, a slight move in one part may affect the situation as a whole. It will reduce the efficiency of this "passive" evaluation and design, and it is difficult to ensure the repeatability of its theoretical methods.

CCD is one kind of hand controller. In the 1990s, with the mouse becoming the main input device and its wide application in the field of personal computers, related literature [2–4] discussed the shape of the mouse and the influence of office space design on upper limb motion constraint and manipulation performance. With the development of sports medicine and biomechanics [5, 14], the ergonomic design of hand controller has a certain theoretical basis. When considering its application scenario, the selection of corresponding design theory will be the theoretical basis to show the certification of this kind of control device. This "active" design and evaluation method will also greatly improve the efficiency of design work.

2 Brief Introduction to the Application of CCD in Civil Aircraft

Cockpit Display System (CDS) has become a human machine interaction platform for modern civil aircraft airborne system. It accesses virtual Graphical User Interface (GUI) components by cursor, realizes the interaction between pilots and many application functions such as navigation map, flight management, integrated surveillance, data link, radio tuning, electronic checklist, and so on, thus realizes platform and modular system architecture, reduces or even abandons independent computing processing and control Unit.

CCD belongs to one of the multifunctional controllers. It generally controls the motion trajectory of the cursor by means of a trackball, joystick, touchpad, etc., in order to focus and control the state of the GUI components and realize the interaction between the pilot and the aircraft system.

Typical interaction between a cursor and GUI components via a CCD is shown in Fig. 1.

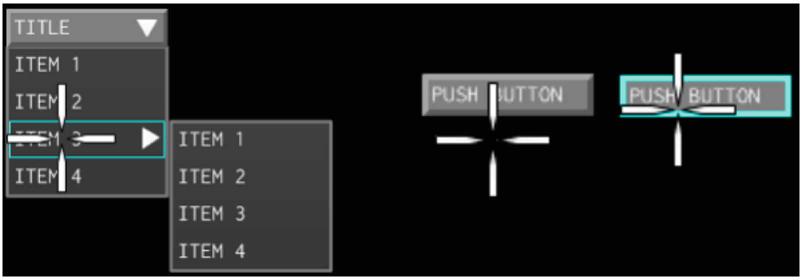


Fig. 1. The interaction between cursor and GUI component

The installation of CCD and ergonomic evaluation scenarios on some aircrafts are shown in Fig. 2.

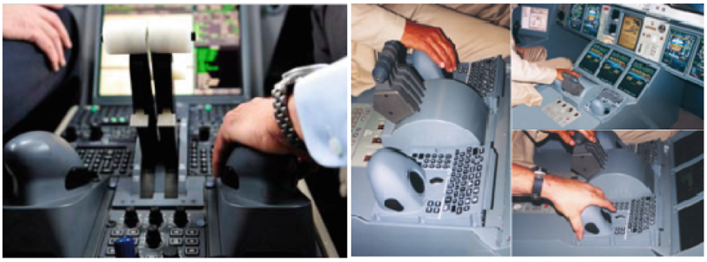


Fig. 2. Examples of CCD installation and evaluation

From the perspective of human machine interaction, the advantages of integrating many functional applications on CDS platform are as follows:

a. The interaction interface concentrates in the pilot's main field of view, which decreases the head movement required for interaction, and the forward-looking attitude helps to maintain the pilot's external situation awareness;
b. Based on GUI interaction specification, the interaction behavior of human machine interface of each system is easy to be unified, which not only reduces the training difficulty and cost, but also lessens the probability of human errors and the number of dedicated control devices.
c. The system can provide flexible and intuitive interactive experience and simplify flight crew's operational action.
d. Facilitate the upgrade and expansion of system functions.

Due to the fact that many system functions rely on CCD for interaction, it is used very frequently in the cockpit [2, 12]. In addition, the CCD usage scenario covers almost all flight phases, including the phase of high work load in the terminal area [7, 15, 17–19] (for example, with frequent switching of communication frequencies, selection of standby flight plans, alteration of approach routes, etc.). Therefore, CCD has been clearly listed as an important cockpit control device in the relevant advisory circular [13], and its importance is equivalent to that of operating equipment such as aviation and throttle control.

Because CCD belongs to hand control equipment, in addition to satisfying the basic requirements of general controller such as two-dimensional layout, lighting, label and accessibility, special requirements such as hand stability, compatible operation in vibration environment and reducing operation fatigue need to be considered.

From the above, the airworthiness compliance criteria of CCD ergonomics design will be significantly different from that of common control devices.

3 Analysis of Relevant Airworthiness Clauses

The provisions related to CCD in 25 regulations of CAAC, FAA and EASA are summarized as shown in Table 1 [8–10]:

In terms of the frequency and workload of operation task, especially installation position, the ergonomic design of the cursor control device will be an important aspect of its airworthiness certification.

Table 1. Airworthiness regulation apply to CCD

Relevant airworthiness clauses	Regulation chapter	Scope of application

Each cockpit control must be located to provide convenient operation
The controls must be located and arranged, with respect to the pilots' seats, so that there is full and unrestricted movement of each control without interference from the cockpit structure or the clothing of the minimum flight crew (established under CS 25.1523) when any member of this flight crew from 1.58 m (5ft 2 in.) to 1.91 m (6ft 3 in.) in height, is seated with the seat belt and shoulder harness (if provided) fastened
	25.777(a)
	25.777(c)
		Cockpit controls

Each pilot compartment and its equipment must allow the minimum flight crew (established under CS 25.1523) to perform their duties without unreasonable concentration or fatigue
Vibration and noise characteristics of cockpit equipment may not interfere with safe operation of the airplane
	25.771(a)
	25.771(e)
		Pilot compartment

Each item of installed equipment must –
(1) Be of a kind and design appropriate to its intended function;
(2) Function properly when installed
	25.1301(a)(1)
	25.1301(a)(4)
		Function and installation

Flight deck controls and information intended for flight crew use mustbe accessible and usable by the flight crew in a manner consistent with the urgency, frequency, and duration of their tasks	25.1302b(2)	Human factor
The minimum flight crew must be establishedso that it is sufficient for safe operation, considering – The accessibility and ease of operation of necessary controls by the appropriate crew member	25.1523 (b)	Minimum flight crew

Considering the functional characteristics of CCD, the related ergonomic provisions can be summarized as follows:

a. Because of the interaction of many functions and its operation frequency [11], the position of CCD should ensure that the pilots in a certain range of height (and arm length) can operate easily, and the operation posture of hands and arms should be conducive to preventing fatigue;
b. The shape and position of CCD should be helpful for pilots to resist the adverse operation effects caused by vibration environment.

For flight deck system controls, the Federal Aviation Administration of the United States issued a special advisory circular (AC) in 2011, namely AC20-175. Considering the convenience of the use of CCD, this AC suggested that the CCD should be placed in or near the pilot's natural hand position, and should be combined with the use of hand stabilization device or arm support device. The aim is to minimize the movement of the hand and arm and to facilitate the operation of pilots, especially when pilots have time pressure.

Compared with the airworthiness clause, AC20-175 further explains the position of the hand and arm when operating the CCD, but the clause also requires that pilots within a certain height (and arm length) range be easy to operate, which is based on anthropometric statistics of Americans in the 1980s, corresponding to 5% of females and 95% of males, respectively. The core of the issue is that the arm length of the crowd in this area is obviously different, and the CCD can only be installed in a fixed position in the cockpit. Then, under normal sitting posture, whether it is possible for the natural hand positions of the above pilot crowd to be concentrated in a limited position range will be the key to prove the airworthiness of CCD ergonomic design.

4 Compliance Strategy of CCD Ergonomic Design

4.1 General Design Principles of Hand Controller

Because of its wide application scenarios and high frequency of use, hand controller has always been one of the hot and difficult points in ergonomics.

In addition to functional design, the design of hand controller also requires comprehensive knowledge of anatomy, anthropometry, and kinesiology. The cumulative effect trauma (such as carpal tunnel syndrome, tenosynovitis, trigger finger, tennis elbow, etc.) caused by improper design has a high incidence in daily life and work.

The basic principles for ergonomic design of hand controller include [14]:

a. Maintain wrist level during operation to avoid compression of adjacent tissues, as shown in Fig. 3;
b. Avoiding repeated finger movements;
c. Inclusive design. That is to say, anthropometric statistics for different races, from 5% female to 95% male and left-handed people are considered.
d. When the carpal canal is in the middle, the grip force is the greatest.

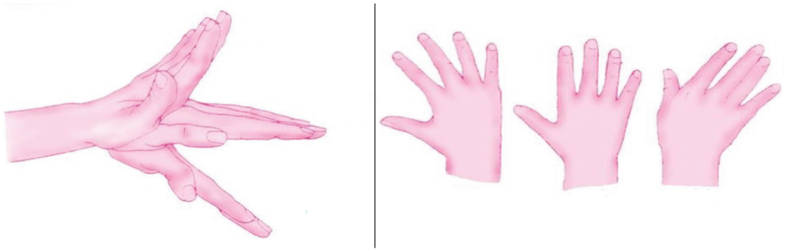


Fig. 3. Gesture of hand

4.2 Consideration of CCD Ergonomic Design

The ergonomic design of CCD is mainly divided into the following aspects:

a. Shape and size of the grip;
b. Mounting position of the grip;
c. Size and position of the trackball or touchpad;
d. Gain and data delay of the trackball or touchpad;
e. Size and position of the cursor selection key.

Obviously, the mounting position of the grip is the decisive factor to support the hand and drive the arm posture through the wrist joint. The shape and size of the grip will determine the hand pattern and the position and size of the trackball and cursor selection key. In addition, the combination of the both is a decisive factor in determining wrist posture. Therefore, the design methods for these two aspects will directly determine the applicable criteria and compliance of airworthiness clauses.

5 CCD Ergonomic Design Method

5.1 Scenario Analysis of CCD Use

The human-machine interaction system which combines display and control generally adopts the principle of "display above, operate below". That is to say, the display interface is generally located above the control device to avoid obscuring display information during operation. Relevant design principles are common in industrial standards. Therefore, the CCD is usually installed on the central pedestal of the cockpit.

As far as the interactive process of virtual control is concerned, the operation tasks of cursor movement and keyboard input are serialized. Therefore, CCD and multifunction keyboard (MKB) should be arranged adjacent to each other. In view of the need of grip support and stabilization of hand, MKB should be placed in front of CCD.

In order to improve the safety of human-machine interaction system operation and aircraft dispatch rate, the operation functions of CCD and MKB should be backed up mutually.

As stated above, the operation of the CCD covers almost all flight phase, so:

a. In the terminal area, pilots should be able to use CCD efficiently when seated at design eye position(DEP) of the cockpit and fastening seat belts and shoulder straps;
b. During the cruise phase, should be able to easily use CCD when flight crew moving the seats backwards and deviates from the cockpit design eye position, and fastening the seat belt, with the lower operating frequency.

5.2 Selection of Appropriate Anthropometric Data

The anthropometric data is the basis of analyzing and designing the position and shape of CCD grip. Ethnic factors should also be taken into account in the selection of anthropometric data. The design data in this paper come from ISO-TR-7250-2 [20] of International Organization for Standardization.

The anthropometric data of typical ethnic groups in the United States, Germany, the Netherlands, Italy, Japan, Korea, Thailand, Kenya and other countries are collected in this standard. It should be noted that there is a big gap between the corresponding percentage data of some ethnic groups and the airworthiness provisions, which should not be considered.

The anthropometric data of Germany, Japan, Korea and Kenya are used in this paper.

5.3 Determine the Operational Posture

According to the above analysis, when the fight crew is seated at the cockpit design eye position, and fastening the seat belt and shoulder strap, if the upper arm is naturally drooping and the palm is put on the grip, it can maintain the wrist at right position. This posture can avoid the muscular tension from the shoulder and arm and minimize the pressure of the shoulder joint and elbow joint. It is more beneficial for pilots to operate CCD comfortably and efficiently for a long time. At this time, with the wrist at the right position and palms in the median position, it is convenient to grasp the CCD grip in adverse environment.

In addition, it should also avoid the occurrence of outer surface of CCD against palms. Improper long time compression will lead to pain and numbness in the arms and hands. Therefore, the muscle group with 1/3 palms from the wrist is chosen as the reference point to support the palm, thus avoiding the nerve intersection of the hand.

To sum up, the reasonable attitude of pilots to operate CCD is shown in Fig. 4.

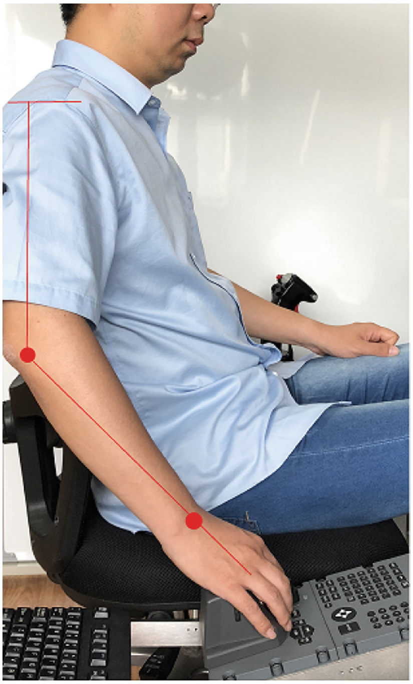


Fig. 4. Proper attitude of handling CCD

5.4 Computer Aided Design and Analysis

The required human anthropometry data include:

a. Height of eye (sitting position);
b. Height of shoulder (sitting position);
c. Distance between shoulder and elbow;
d. Distance between elbow and wrist;
e. Palm length;
f. Width between two acromia.

Based on the above data, the geometric parameters of a person operating CCD in a reasonable manner under the sitting position are sorted out. According to the movement ability of the human arm, if the upper arm is kept naturally drooping, in theory the forearm will be in front of the human body with elbow joint as the center, drawing about 1/4 of the spherical surface, in this problem only take the second half of the spherical surface, that is, 1/8 of the spherical surface as a research reference. When the eye coordinates corresponding to the sphere are aligned with the design eye position of the cockpit, the envelope of operating CCD can be obtained by each race in a reasonable attitude, as shown in Fig. 5.

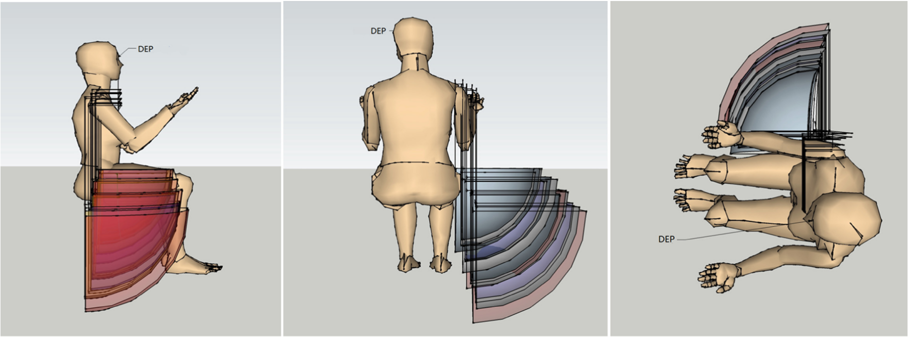


Fig. 5. Proper envelope of operating CCD

On the side of the central pedestal which is close to the human body, the intersection of the vertical plane along the heading and the envelope surface will be obtained as shown in Fig. 6.

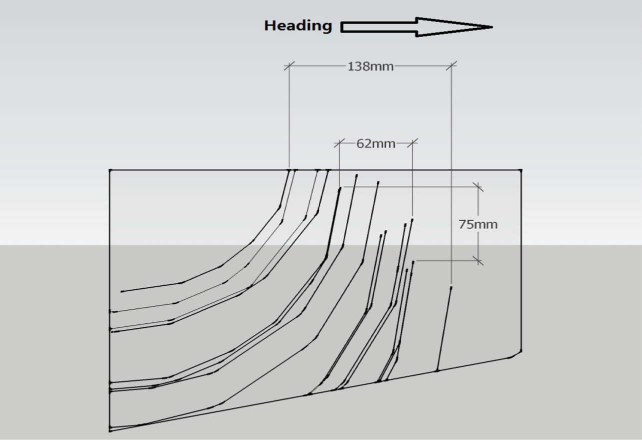


Fig. 6. Intersection line of envelope on the vertical plane

The arcs on the vertical plane in Fig. 6 show the distribution of all palm reference points of several ethnic groups when they operate CCD reasonably. On the far left is 5% Korean women and on the far right is 95% Kenyan men. Considering that 50% of the population, that is, the majority of the population should be comfortable to operate, it is not difficult to find that when focusing on the analysis of the crossings of 50% of the population, these crossings are relatively concentrated, and can even contain almost 95% of German men.

After measuring and analyzing the computer model, the support reference points of the palm in the operation of CCD are approximately distributed in an area of about 62 mm in length and 75 mm in height along the heading. The size of this area is much smaller than the difference in height (arm length) in anthropometric data.

If it is considered that in cruise phase, the pilots can easily use CCD when moving their seats backwards, it is suggested to select the reference points which are a little back in that area so as to be compatible with the pilots' operation at that time.

6 Conclusions

In this paper, the airworthiness clauses and advisory circulars applicable to CCD are analyzed. According to the daily use scenarios of CCD, based on the design principle of hand controller, the human anthropometry data in ISO standard is extracted extensively. The posture with reasonable operation on CCD is modeled by computer aided design technology, and the following conclusions are drawn:

a. In the case of sitting position, the size of the body has a limited impact on the location of CCD installation;
b. In theory, there is a limited position range, which can satisfy the pilot to operate CCD conveniently and efficiently in all flight phases, so as to comply with the requirements of airworthiness clauses;

c. The calculation method and results in this paper can be used as one of the methods and criteria for airworthiness compliance of CCD ergonomic design.

References

1. Su, R.E., Xue, H.J., Song, B.F.: Ergonomic virtual assessment for cockpit layout of civil aircraft. Syst. Eng. - Theory Pract. **29**(1), 186–191 (2009). (in Chinese)
2. Thomas, J.A., et al.: A conceptual model for work-related neck and upper-limb musculoskeletal disorders. Scand. J. Work Environ. Health **19**(2), 73–84 (1993)
3. Armstrong, T.J., Martin, B.J., Arbo, A., Rempel, D.M., Johnson, P.W.: Mouse input devices and work-related upper limb disorders. WWDU **3**(C), 20–21 (1994)
4. Richard, P., Robin, C.: Design criteria of an ergonomic mouse computer input device. SAGE J. **39**(5), 369–373 (1995)
5. Guo, X., Fan, Y.B., Li, Z.M.: Carpal tunnel syndrome and bio-mechanical studies on CTS. Adv. Mech. **35**(4), 472–480 (2005). (in Chinese)
6. Michelle, Y., Cathy, S., Young, J., Colleen, D.: Human factors considerations in the design and evaluation of flight deck displays and controls. DOT/FAA/TC-16/56 (2016)
7. Federal Aviation Administration. TSO-C165A Electronic Flight Instrument. FAA (2015)
8. Civil Aviation Administration of China. CCAR-25-R4 Airworthiness Standards: Transport Category Airplanes. CAAC (2011)
9. Federal Aviation Administration. FAR Part25 Airworthiness Standards: Transport Category Airplanes. FAA (2018)
10. European Aviation Safety Agency. CS-25 Certification Specifications for Large Aeroplanes. EASA (2018)
11. ANM-11. AC25.1302-1 Installed Systems and Equipment for Use by the Flightcrew. FAA (2013)
12. ANM-11. AC25-11B Electronic Flight Displays. FAA (2014)
13. AIR-120. AC20-175 Controls for Flight Deck Systems. FAA (2011)
14. Mark, S., Ernest, J.: Human Factors in Engineering and Design (Trans. by Yu, RF, Lu, L), 7th edn. Tsing Hua University Press, Beijing (2009). (in Chinese)
15. AIR-130. AC20-138D Airworthiness Approval of Positioning and Navigation Systems (Including Change 2). FAA (2014)
16. Committee S-7. SAE ARP4102 Flight Deck Display Panels, Controls, and Displays. SAE International, Warrendale, PA (2007)
17. Radio Technical Commission for Aeronautics. DO-257A Minimum Operational Performance Standards for the Depiction of Navigation Information on Electronic Maps. RTCA (2003)
18. Radio Technical Commission for Aeronautics. Minimum Operational Performance Standards for Required Navigation Performance for Area Navigation. RTCA (2015)
19. Federal Aviation Administration. TSO-C115D Flight Management System Using Multi-Sensor Inputs. FAA (2013)
20. International Organization for Standardization. ISO/TR 7250-2 Basic Human Body Measurements for Technological Design-Part 2: Statistical Summaries of Body Measurements from National Populations. ISO (2010)

Study on Evaluation of Airline Pilot's Flight Violation Behaviors and Psychological Risk

Jingyi Zhang and Lei Wang(✉)

Flight Technology College, Civil Aviation University of China,
Tianjin 300300, China
wanglei0564@hotmail.com

Abstract. This study aimed to evaluate the flight violation behaviors of airline pilots and examine the relationship between violation behavior and risk psychology based on Quick Access Recorder (QAR) data and surveys. Flight violation evaluation indexes were selected from airlines' Flight Operations Quality Assurance (FOQA) items. Then, an evaluation standard for violation behaviors was determined by investigating airlines, and a violation evaluation model for pilots was established. To examine the model's reasonableness and explore the relationship between violation behavior and risk psychology, flight QAR data were analyzed, and pilot's risk psychology characteristics were investigated by using psychological scales. In the case study, correlation analysis showed that landing vertical overload—a key factor in landing safety—was significantly negatively correlated with risk tolerance and significantly positively correlated with risk perception. Significant correlations among the violation indexes indicated interrelationships among the violation behaviors. This evaluation method can be applied to airlines' FOQA to effectively and efficiently identify and control pilot's violation behaviors. These findings are expected to provide a support for improving aviation safety.

Keywords: Violation behavior · Risk psychology · Flight data · Landing safety

1 Introduction

In the past two decades, nearly 75% of civil aviation accidents have been caused by human factors. Pilot's flight safety operations affect human error and safety in civil aviation. In its 2017 Annual Safety Report, the International Air Transport Association noted that 64% of flight accidents caused by flight crew errors could be attributable to manual handling and flight control by pilots [1]. Therefore, in the process of daily operations and management, airlines need to monitor and analyze Quick Access Recorder (QAR) exceedance warnings triggered by pilots to record their flight safety operations. Meanwhile, according to Boeing, while the takeoff, initial climb, final approach, and landing phases account for only 6% of total flight time, 61% of accidents and fatalities occur in these phases, and more than 70% are caused by flight crew errors [2]. Studies have indicated that pilot's ignorance of regulatory frameworks is the main cause of accidents during the final approach and landing phases [3]. In this regard, the

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 342–352, 2019.
https://doi.org/10.1007/978-3-030-22507-0_27

IATA noted that 50% of fatal accidents caused by flight crew errors pertain to Standard Operating Procedure (SOP) adherence and cross-verification issues [1]. Hence, studying pilot's violation behaviors is essential for civil aviation safety.

Previous studies have shown that pilots involved in aviation accidents are more likely to break with regulatory frameworks than those not involved in accidents [4–7]. Rebok et al. selected 3,000 pilots aged 45–54 as samples and tracked their violations for 11 years. The results showed that the risk of violation was negatively correlated with flight experience, and there was a significant positive correlation between age and violation [8]. English and Branaghan constructed a new classification based on pilot's violation intention, grouping the reasons for pilot violations into four categories: improvement, malevolent, indolent, and hedonic [9]. Luo suggested that most behavioral mistakes have to do with psychological characteristics arising from interactions among the crew, the machine, and environmental factors [10]. Liang also focused on psychological factors, investigating common unhealthy psychological factors affecting violations from a micro perspective [11].

While airlines monitor and manage cockpit exceedance by pilots as part of their daily operations, violation behaviors tend to be ignored until there is severe exceedance. An exceedance event is an unsafe event in which any QAR monitoring parameter exceeds the flight operations quality assurance (FOQA) standard, which specifically focuses on the collection and analysis system of flight data in daily flight [12, 13] and is reported by FOQA software. Exceedance event risk management based on QAR data comprises the core of airlines' FOQA. Similar to the idea of big data, QAR flight data can be used to analyze and evaluate pilot's operation levels and exceedance behaviors. Wang et al. investigated the evaluation of flight operation risk using QAR data. This included a study of the flaring operational characteristics of long landing and hard landing events, specifically focusing on the effect of flaring operation on landing performance [14–18]. Overall, even though research has been conducted on using QAR data in the detection, diagnosis, and prediction of exceedance events, few studies have investigated the relationship between risk-related personality traits and exceedance behaviors.

Therefore, the current study screened exceedance events related to flight violation firstly. Then, an evaluation method for pilot violation was constructed whereby violation level could be quantitatively evaluated based on flight QAR data. Meanwhile, the scale of measuring pilot's psychological risk was also introduced and implemented. Methods and results can provide technical support for flight safety management and improve airlines' daily monitoring and safety management of pilot violations.

2 Evaluation on Airline Pilot's Flight Violation Behaviors Based on Flight QAR Data

2.1 Flight QAR Data Acquisition

The QAR is a system that includes equipment for recording data in the air and a software station on the ground for storing and analyzing the data. QAR can record all kinds of aircraft parameters, pilot operation parameters, environmental features, and alarm information during a flight. When a flight parameter exceeds the prescriptive normal range, it is called a QAR exceedance event. While most exceedance events do not produce severe results, they can increase the likelihood of an accident, potentially harming aircraft and even passengers. Based on related operational rules and regulations, commercial airlines always use flight data (such as QAR data) to monitor and analyze the aircraft and the pilot's operational performance in flight. FOQA monitoring standards are developed based on aircraft design principles and flight environments, and are combined with the operation requirements of different airlines. In this study, FOQA standards for the Boeing 737-800 (B737-800) were selected as the research foundation, and exceedance events and flight QAR data were collected to analyze pilot's violation behaviors and establish a violation operation evaluation model. A program based on VBA (Visual Basic for Applications) was written and applied to minimize file volume and mine target information from the massive QAR data.

2.2 Investigation of Violation Event in Airlines

Flight violation events in this study were defined as those exceedance events that were mainly caused by pilot's subjective intentions. This kind of exceedance events with violation were selected through discussion and analysis of the causes as well as the related operating manuals. To screen for reasonable and effective indicators, several airlines were investigated by communicating with expert pilots, flight instructors, and FOQA professionals. According to the task characteristics, exceedance events were classified preliminarily by different flight phases.

For the B737-800 aircraft, there are 82 indicators for QAR monitoring standards in the airline selected for the current research. Through discussion and investigation, 38 violation behavior items were eventually selected. Figure 1 shows the violation events involved in each flight phase.

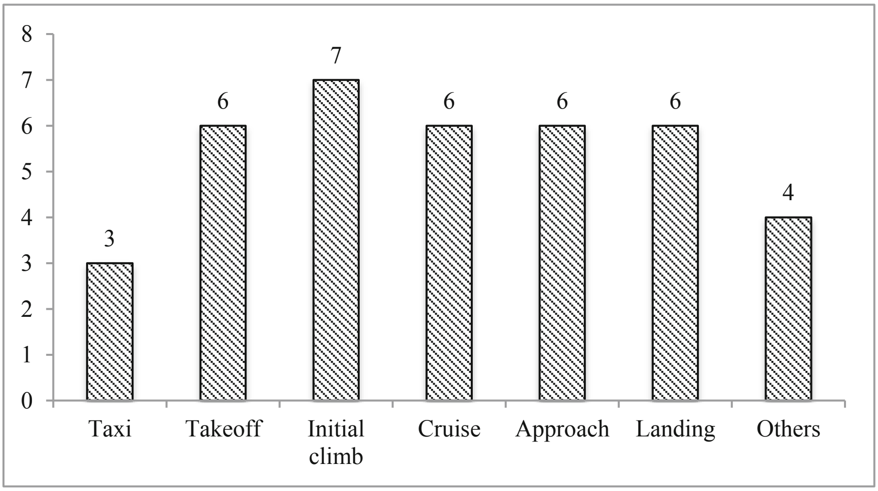


Fig. 1. Number of each violation type in different phases of flight

Violation types occurring at the selected airline from year of 2014 to 2017 were collected and calculated, and then sorted by frequency of occurrence. The results are shown in Table 1.

Table 1. Statistics for violation frequency from 2014 to 2017

Mild exceedance		Severe exceedance	
Event	Frequency	Event	Frequency
Landing vertical overload	281	Ground Proximity Warning System (GPWS) warning	26
Landing gear up late	19	Landing vertical overload	5
Cornering taxiing overspeed	16	Landing gear up late	4
Landing flaps in position late	12	Landing flaps in position late	2
Straight taxiing overspeed	5	Exceeding tire limit speed	1
Landing gear down late	5	Cornering taxiing overspeed	1

2.3 Evaluation Model of Airline Pilot's Violation Behaviors

Classification and Selection of Evaluation Indexes. The occurrence frequency of violation events from 2014 to 2017 was taken as an optimizing factor for the violation evaluation indexes. If a violation event had not occurred in nearly four years, the index was deleted; if a violation event had occurred within four years, the index could be retained. Table 1 shows that in the past four years, there were eight violation items triggered by pilots: straight taxiing overspeed, cornering taxiing overspeed, exceeding tire limit speed, landing gear up late, Ground Proximity Warning System (GPWS) warning, landing gear down late, landing flaps in position late, and landing vertical overload. However, since the exceeding tire limit speed event only occurred once, the index was deleted. Hence, the final evaluation indexes were as follows: E1, straight taxiing overspeed; E2, cornering taxiing overspeed; E3, landing gear up late; E4, GPWS warning; E5, landing gear down late; E6, landing flaps in position late; and E7, landing vertical overload.

In terms of object characteristics recognized by the human brain, perception can be divided into three categories: space perception, temporal perception, and motion perception [19]. Furthermore, airline pilots must also accurately judge the position and motion state of the aircraft. Also, given the close association between temporary perception and space perception, violation events can be classified into two categories: (1) caused by space perception errors and (2) caused by motion perception errors.

Using this method, the seven violation event items were classified. Table 2 shows the results, indicating that these seven indexes of violation evaluation can be covered.

Table 2. Classification of violation indexes

Space perception errors	Landing gear up late
	Landing gear down late
	Landing flaps in position late
	GPWS warning
Motion perception errors	Straight taxiing overspeed
	Cornering taxiing overspeed
	Landing vertical overload
	GPWS warning

Determination of Evaluation Standard. Further analysis of QAR data showed that although some parameters of flight operation did not exceed FOQA standards, they were very close to the critical value and tended toward exceedance. This part of the data is not marked in the monitoring system and is thus often overlooked in airlines' daily management. Although events tending toward exceedance are not recorded by FOQA, there is great potential for triggering exceedance in more complex situations. Therefore, in addition to exceedance events recorded by the FOQA software, violation tendencies should also be considered when determining the evaluation standard for violation behaviors. QAR flight data should be fully used to guarantee flight safety.

The evaluation standard for pilot's violation behaviors was based on the monitoring items and standards for severe and mild exceedance for the B737-800 at the selected airline. Tendency toward violation behavior was graded by the percentile method as the evaluation standard for each index. QAR flight data were collected and extracted corresponding to each index and then sorted from smallest to largest. The accumulated percentage was calculated using SPSS software. Due to different types of index data, some exceedance standards were on the large side and some on the small side; thus, 70 or 30 percentiles were selected as the classification standard for tendency violation. The percentile formula is shown as follows:

$$P_p = L_b + \frac{\frac{P}{100} \times N - F_b}{f} \times i \tag{1}$$

where P_p is the percentile of P, L_b is the exact lower limit of the group in which the percentile is located, f is the frequency of the group in which the percentile is located, F_b is the frequency sum for each group that is less than L_b, N is the total frequency, and i is the class interval.

If the QAR flight data exceeded the severe exceedance standard, it was scored as 30 points. Similarly, 20 points were allocated for mild exceedance and 10 points for tendency exceedance. Finally, the evaluation standard for airline pilot's violation behaviors is as shown in Table 3.

Table 3. Evaluation standard for airline pilot's violation behaviors

Violation event	Severe exceedance		Mild exceedance		Tendency exceedance	
	Exceedance standard	Scoring	Exceedance standard	Scoring	Exceedance standard	Scoring
E1 Straight taxiing overspeed	$\geq$ 40Kts	30	$\geq$ 30Kts	20	<23Kts $\geq$ 23Kts	0 10
E2 Cornering taxiing overspeed	$\geq$ 18Kts	30	$\geq$ 14Kts	20	<12Kts $\geq$ 12Kts	0 10
E3 Landing gear up late	$\geq$ 500Ft	30	$\geq$ 300Ft	20	<104.4Ft $\geq$ 104.4Ft	0 10
E4 GPWS warning	Detected	30	–	–	–	–
E5 Landing gear down late	$\leq$ 1300Ft	30	$\leq$ 1500Ft	20	>1991.6Ft $\leq$ 1991.6Ft	0 10
E6 Landing flaps in position late	$\leq$ 1000Ft	30	$\leq$ 1200Ft	20	>1852.6Ft $\leq$ 1852.6Ft	0 10
E7 Landing vertical overload	$\geq$ 1.89 g	30	$\geq$ 1.68 g	20	<1.455 g $\geq$ 1.455 g	0 10

Calculation of Index Weight. The entropy weight method, an objective weighting method, was used to calculate the weight of the evaluation indexes. The entropy value reflects the disorder degree of information. The smaller the value, the smaller the disorder degree of the system. For the discrete degree of the indexes, the larger the value, the greater the discrete degree; that is, the greater the effect on the violation evaluation system. The steps for weight calculation are shown as follows:

(1) Standardization of QAR flight data

$$X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\} \tag{2}$$

$$Y_i = \{y_{i1}, y_{i2}, \ldots, y_{in}\} \tag{3}$$

$$y_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \tag{4}$$

where X_i is the original indexes given for violation evaluation, Y_i is the standardized indexes, i = 1,2,…,7, and j = 1,2,…,n.

(2) Calculation of entropy:

$$D_j = -\ln(n)^{-1} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \tag{5}$$

where D_i is the entropy of index i, $p_{ij} = Y_{ij} \Big/ \sum_{i=1}^{n} Y_{ij}$, if P_{ij} = 0, $\lim_{p_{ij}=0} p_{ij} \ln p_{ij} = 0$

(3) Calculation of weight:

$$W_i = \frac{1 - D_i}{7 - \sum D_i} \tag{6}$$

where W_i is the weight of index i.

Evaluation Model for Violation Behavior. 348 sets of QAR flight data corresponding to 27 pilots were collected. For each index, the necessary parameters were extracted using MATLAB. Then, the extracted parameters were calculated according to the evaluation standard for each index. The calculating formula is shown as follows:

$$X_i = \frac{\sum_{j}^{n} Z_j}{n} \tag{7}$$

where X_i is the pilot's score of index i, n is the number of flights, and Z_i is the score of flight j.

$$Z_j = \left\{ \begin{array}{l} 0, \, Normal \\ 10, \, TendencyExceedance \\ 20, \, MildExceedence \\ 30, \, ServerExceedence \end{array} \right\}$$

Through the comprehensive quantification of each violation index, the violation evaluation model was established:

$$L_k = \sum W_i X_i \tag{8}$$

where L_k is the comprehensive violation evaluation result for each pilot, W_i is the weight of index i, X_i is the score of index i, i = 1,2,…, 7, and k = 1,2,…, 26.

3 Measurement on Pilot's Psychological Risk

3.1 The Scale of Risk Psychology

On the basis of previous studies [20], three risk psychological characteristics were selected as the evaluation indicators of risk psychology. To measure pilot's risk psychology characteristics during flight operation, a scale was established by modifying and translating the risk tolerance, risk perception, and hazardous attitude scales. The test results showed that the scales had good reliability and validity.

Scale Structure. The risk tolerance scale for pilots comprises 17 kinds of flight scenarios, established according to Hunter [21] and Ji et al. [20]. The risk tolerance score is measured in five grades: 5, pilot is extraordinarily willing to accept, or agree with, the flight scenarios given on the scale; 4, pilot is willing to accept, or agree with, the flight scenarios given on the scale; 3, pilot is not sure or indifferent to the flight scenarios given on the scale; 2, pilot is reluctant to accept, or agree with, the flight scenarios given on the scale; and 1, pilot is very reluctant to accept, or agree with, the flight scenarios given on the scale.

The risk perception scale for pilots consists of 26 kinds of flight scenarios. It has been widely applied in research since its creation by Hunter [21]. The grades of risk scenarios listed in the scale range from 0 to 100.

The hazardous attitude scale for pilots comprises 24 kinds of behavior that are closely related to modern airline activities. It is a 5-point Likert scale: 5, pilot is extraordinarily willing to accept, or agree with, the flight situation given on the scale; 4, pilot is willing to accept, or agree with, the flight situation given on the scale; 3, pilot is not sure or indifferent to the flight situation given on the scale; 2, pilot is reluctant to accept, or agree with, the flight situation given on the scale; and 1, pilot is very reluctant to accept, or agree with, the flight situation given on the scale.

The final score for each scale is the average score for all of the topics. The higher the final score, the higher the level of risk psychology characteristics.

Scale Implementation. The average age of the 27 male pilots who participated in the flight data acquisition and questionnaire survey was 29.22 years. The average total flight hours for three years was 2,636.81 h. Table 4 shows the basic statistical data of the subjects.

Table 4. Basic statistical data of the subjects

	Hierarchy	Number	Proportion
Age	21–25	4	15.38%
	26–30	17	65.38%
	31–35	3	11.54%
	36–41	3	7.69%
Technical grade	Instructor	3	7.69%
	Captain	10	38.46%
	First officer	14	53.85%
Flight hours in three years (2015–2017)	1000–1500	1	3.85%
	1501–2000	1	3.85%
	2001–2500	4	11.54%
	2501–3000	21	80.77%

3.2 Correlation Between Risk Psychology Characteristics and Airline Pilot's Violations

Finally a case was given by using the evaluation method and survey results. The Pearson correlations between the scores of risk psychology characteristics and flight violation behaviors in landing was analyzed. Since the 27 pilots scored the same on the GPWS warning index, this item was excluded from the correlation analysis. Landing vertical overload was significantly negatively correlated with risk tolerance *($r = -0.474$, $p < 0.05$)* and significantly positively correlated with risk perception

$(r = 0.585, p < 0.05)$. The negative relationship between cornering taxiing overspeed and risk perception was significant *$(r = -0.468, p < 0.05)$*. Furthermore, there was a significantly negative correlation between risk tolerance and risk perception *$(r = -0.547, p < 0.05)$*. For violation behaviors, landing gear up late showed a significantly negative correlation with straight taxiing overspeed *$(r = -0.444, p < 0.05)$*. Further, landing gear down late showed significantly positive correlations with landing gear up late *$(r = 0.441, p < 0.05)$* and landing flaps in position late *$(r = -0.686, p < 0.05)$*.

4 Discussion

4.1 Theoretical Model for Airline Pilot's Violations

An evaluation model was established based on the investigation and statistical analysis. The correlations among violation indexes—such as landing gear down late, landing gear up late, and landing flaps in position late—indicated close associations among certain flight operation violation behaviors, indicating that it is reasonable to classify these seven violation behavior items based on the perspective of perception. However, the interrelationships were not entirely consistent with the basis of classification—that is, the different characteristics of perceptual objects. For motion perception errors, landing gear down late significantly positively correlated with landing gear up late and landing flaps in position late, though there were no significant interrelationships among the four violation indexes of space perception errors. Instead, straight taxiing overspeed showed a significantly negative correlation with landing gear up late, which might be attributable to the close association between space perception and motion perception. Previous studies have shown that an interrelationship exists between space perception and motion perception. That relationship still needs further experimental exploration; thus, the model needs to be further optimized. This could also be attributable to sample size restrictions; thus, the model should be verified by further case studies in follow-up research.

4.2 Effect of Risk Psychology Characteristics on Landing Operation

Contrary to previous findings, correlation analysis showed that landing vertical overload was significantly negatively correlated with risk tolerance and significantly positively correlated with risk perception [22–27]. Relevant surveys and investigations indicated that, today, airlines emphasize hard landing monitoring and adopt more severe punishment measures for hard landing exceedance events than for other events. Therefore, pilots tend to deliberately prolong flare time and touchdown distance to minimize the landing vertical load. Wang et al. found a significant correlation between touchdown distance and average landing vertical load [14], indicating that flights with a longer touchdown distance generally have a lighter vertical load in landing. However, a long flare time or touchdown distance can lead to another exceedance, long landing, which can trigger a more severe accident—overshooting the runway—which can cause serious economic losses and even casualties. So, when pilots lengthen flare time and extend touchdown distance to avoid a hard landing, they increase the risk of running off the runway. With increased flare time, the touchdown point becomes more distant, which can increase pilot's psychological pressure. In this case, pilots have higher risk tolerance. Meanwhile, pilots with higher risk perception would prefer a shorter flare time due to fears of running off the runway. In that case, the aircraft would touchdown in a relatively shorter range, which could produce a larger vertical load than would be the case with flights performed by pilots with lower risk perception.

5 Conclusion

In the current study, 38 flight violation event items mainly caused by subjective risk taking by pilots were identified. Furthermore, an evaluation method for pilot's flight violation behaviors was developed, which can be applied to airlines' FOQA to effectively and efficiently identify and control pilot's violation behaviors.

The correlation analysis between violation behaviors and risk psychology indicated that pilots with high severe exceedance rates had higher hazardous attitude scores than pilots with high scores for landing vertical overload violations, who generally possessed low levels of risk tolerance and risk perception. This could be attributable to strict systems of punishment and safety cultures in airlines. These findings are expected to provide new ways for airlines to establish effective management systems and positive safety cultures, thereby improving aviation safety.

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. International Air Transport Association.: IATA Safety Report 2017 (2018)
2. Boeing Commercial Airplanes.: Statistical Summary of Commercial Jet Airplane Accidents Worldwide Operations 1959–2015 (2016)
3. Civil Aviation Administration of China.: Annual report on aviation safety in China 2017. Civil Aviation Administration of China (2018)
4. Burg, A.: Traffic violations in relation to driver characteristics and accident frequency (Rep. No. 74–55). Los Angeles: University of California Los Angeles, Institute of Transportation and Traffic Engineering (1975)
5. Civil Aviation Administration of China: Implementation and Management of Flight Operation Quality Assurance. Advisory Circular: 121/135–FS–2012–45. Beijing, China (2012)
6. Chen, W., Cooper, P.J., Pinili, M.: Driver accident risk in relation to the penalty point system in British Columbia. J. Saf. Res. **26**(1), 9–18 (1995)
7. Lourens, P.F., Vissers, J.A.M.M., Jessurun, M.: Annual mileage, driving violations, and accident involvement in relation to drivers' sex, age, and level of education. Accid. Anal. Prev. **31**(5), 593–597 (1999)

8. Rebok, G.W., Qiang, Y., Baker, S.P., Mccarthy, M.L., Li, G.: Age, flight experience, and violation risk in mature commuter and air taxi pilots. Int. J. Aviat. Psychol. **15**(4), 363–374 (2005). https://doi.org/10.1207/s15327108ijap1504_4
9. English, D., Branaghan, R.J.: An empirically derived taxonomy of pilot violation behavior. Saf. Sci. **50**(2), 199–209 (2015)
10. Luo, F.: The Psychological analysis on the behavior mistakes of the civil aviation crews. J. Wuhan Univ. Technol. (4), 687–705 (2002)
11. Liang, L.: An analysis of violation behavior and psychological factors. China Saf. Sci. J. **8** (2), 4–8 (2006)
12. Civil Aviation Administration of China: Implementation and Management of Flight Operation Quality Assurance. Advisory Circular: 121/135–FS–2012–45. Beijing, China (2012)
13. Yan, H.: Building model for relationship between road traffic accident and drivers' psychological quality. China Saf. Sci. J. **26**(2), 13–17 (2016)
14. Wang, L., Ren, Y., Wu, C.: Effects of flare operation on landing safety: a study based on ANOVA of real flight data. Saf. Sci. **102**, 14–25 (2018)
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**, 88–92 (2014)
16. 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
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.: An analysis of flight Quick Access Recorder (QAR) data and its applications in preventing landing incidents. Reliab. Eng. Syst. Saf. **127**, 86–96 (2014)
19. Peng, D.: Common Psychology, 4th edn. Beijing Normal University Publishing Group, Beijing (2015)
20. Ji, M., You, X.Q., Lan, J., Yang, S.: The impact of risk tolerance, risk perception and hazardous attitude on safety operation among airline pilots in China. Saf. Sci. **49**(10), 1412–1420 (2011)
21. Hunter, D.R.: Risk perception and risk tolerance in aircraft pilots (2002)
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 US Army aviators (2009)
24. O'Hare, D.: Pilots' perception of risks and hazards in general aviation. Aviat. Space, Environ. Med. (1990). https://doi.org/10.1080/15298669091369853
25. Platenius, P.H., Wilde, G.J.: Personal characteristics related to accident histories of Canadian pilots. Aviat. Space Environ. Med. **60**(1), 42 (1989)
26. Wiggins, M., Connan, N., Morris, C.: Weather-related decision making and self-perception amongst pilots. Appl. Aviat. Psychol.- Achiev. Chang. Chall. 193–200 (1996)
27. Ji, M., Liu, Z., Yang, S., Bao, X., 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)

Study of NASA-TLX and Eye Blink Rates Both in Flight Simulator and Flight Test

Yiyuan Zheng(✉) and Yuwen Jie

Shanghai Aircraft Airworthiness Certification Center of CAAC, Shanghai, People's Republic of China
zhengyiyuan@saacc.org.cn

Abstract. In order to determine the minimum flight crew number and to show compliance with aircraft airworthiness regulations of CS25.1523, the workload of flight crew should be measured in various fight scenarios both in flight simulator and in flight test. However, the complexity, environment and safety consideration of flight test requires flight crew to take more responsibility and more careful with decisions and actions with higher stress, and it may be inappropriate to carry out the flight test in a high-risk abnormal situation. Therefore, it is necessary to assess workload measures in a simulator to predict in-flight behavior.

In this research, NASA-TLX and eye blinks rate were compared, both in flight simulator and in flight test in three flight scenarios, including Standard Instrument Departure, Manual Departure, and Standard Instrument Approach. This study were carried out in a CRJ-200 full - flight simulator and an aircraft, and a total of nine pilots were participated in.

According to the results, both flight scenarios and environments had the significant influence on NASA-TLX. However, eye blinks rate only manifested significant differences in flight environment. Furthermore, the relation between NASA-TLX and eye blinks rate are weak between simulator and flight test. Therefore, in order to reduce the quantity and risk of compliance demonstrating flight test, it is necessary to figure out more significant psychophysiological measurements.

Keywords: Airworthiness · Workload · Flight test · Flight simulator

1 Introduction

For commercial aircraft, airworthiness is certification and supervision on the design, manufacture, implementation and maintenance of the aircraft according to the airworthiness regulations and materials on behalf of public [1]. The aim of airworthiness is to ensure the aircraft could achieve the safety level that the regulations required. Typically, the design of commercial aircraft should comply with Certification Specifications for Large Aeroplanes CS-25, which is issued by European Aviation Safety Agency [2].

Human factors is the most important factors that could threaten aviation safety. According to the statistics, over 70% flight accidents were attributed to human factors [3]. There are several airworthiness regulations that concerning human factors in CS-25. Among them, CS25.1523-Minimum Flight Crew, is one of the most important regulations

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 353–360, 2019.
https://doi.org/10.1007/978-3-030-22507-0_28

which stipulates the determination of the number of flight crew should base on the workload on individual crew members. In other words, in order to show the compliance with CS25.1523, the workload of each flight crew member should be measured. Furthermore, the recommended means of compliance includes simulator test and flight test.

Typically, the traditional workload measurements for flight crew consist of four types: timeline analysis, task performance measures, subjective rating scale measures and psychophysiological measures [4]. Timeline analysis could be used as an analytic tool in order to make a priori predictions regarding the task demands imposed on the crew [5]. It based on micro-motion techniques and borrowed from industrial engineering, computes workload as a ratio of time required to complete necessary tasks as a fraction of time available. In several aircraft types design, Boeing Commercial Airplane used timeline analysis technique in simulator studies [6]. Task performance measures can be classified into two major types: primary task measures and secondary task measures [7]. Normally, performance of the primary task will always be of interest as its generalization be central to the study. Speed, accuracy, response times, and error rates are often used to assess primary task performance [8]. Bliss and Dunn supported the hypotheses that increasing primary task and alarm task workload degraded alarm response performance [9]. The secondary task technique assumes that operators are given an additional information processing task to perform in conjunction with the task of interest. The rationale underlying the use of secondary tasks is that by applying an extra load which produces a total information processing demand that exceeds the operator's capacity, workload can be measured by observing the difference between single task and dual task performances [10]. Wester et al. examined the impact of secondary task performance, an auditory oddball task, on a primary lane keeping driving task [11]. By studying the impact of simultaneous information conflicts, from multiple secondary in-vehicle tasks, on the primary task of driving, Lansdown and Brook-Carter suggested overloading the visual channel would result in performance decrements [12]. Subjective rating scale measures assume that an increased power expense is linked to the perceived effort and can be appropriately assessed by individuals. NASA-TLX, Bedford scale, and Modified Cooper-Harper scale are most popular ones. Schnell et al. evaluated Synthetic vision information systems in flight deck by using NASA-TLX [13]. The pilot workload, which was assessed through Bedford scale, resulting from a range of wind-over-deck conditions have been used to develop the Ship-Helicopter Operating Limits for a Lynx-like helicopter and the SFS2 [14]. Physiological measures use the physical reactions of the body to objectively measure the amount of mental work a person is experiencing. It would seem an objective measurement would be the most exact and therefore the best way to find workload because it does not require a direct response from the person, unlike subjective measures [15]. In physiological areas, eye activity and cardiac activity are the most research focuses on. Heart rate measurement is considered the most common and reliable measure of workload. Generally, heart rate increases as workload increases [16]. Moreover, eye activity, including pupil dimension and eye blink rate could also indicate the workload. Normally, pupil diameter is found to increase with increasing mental workload, and eye blinks rate decrease with increasing workload [17].

Since flight test may include high or medium risk scenarios, it is necessary to select the appropriate workload measurement which would not interfere with flight crew operation. Therefore, in order to determining the desirable workload measurement in simulator and flight test, subjective rating scale measures and physiological measures, including NASA-TLX and eye blinks rate, were analyzed in this study. Furthermore, 9 pilots composed 6 flight crews were participated in this test which contained three flight scenarios: Standard Instrument Departure (SID), Manual Departure (MD), and Standard Instrument Approach (SIA).

2 Method

2.1 Subjects

Nine Chinese male pilots ranging in age from 30 to 50 (Mean = 41.3 ± 5.23) were invited to participate in this experiment. These pilots were either commercial airline pilots or flight instructors from China Eastern Airlines. Simultaneously, they had all been recruited as captains or co-captains for some types of aircrafts (5 for B737, 4 for B747). Furthermore, these pilots were paired into six flight crews. Among them, three pilots were assigned with different flight responsibilities in different crews involved, i.e., as Pilot Flying in one crew and as Pilot Monitoring in the other. Before the experiment, all subjects signed the consent form, which was approved by the Institutional Review Board of Shanghai Jiao Tong University.

2.2 Apparatus

The experiment was carried out in a CRJ-200 full - flight simulator. It is a qualified flight simulator (level D). All the configurations in the flight simulator are identical with the real aircraft. Simultaneously, the flight test was conducted in a real CRJ-200 aircraft, shown as in Fig. 1.

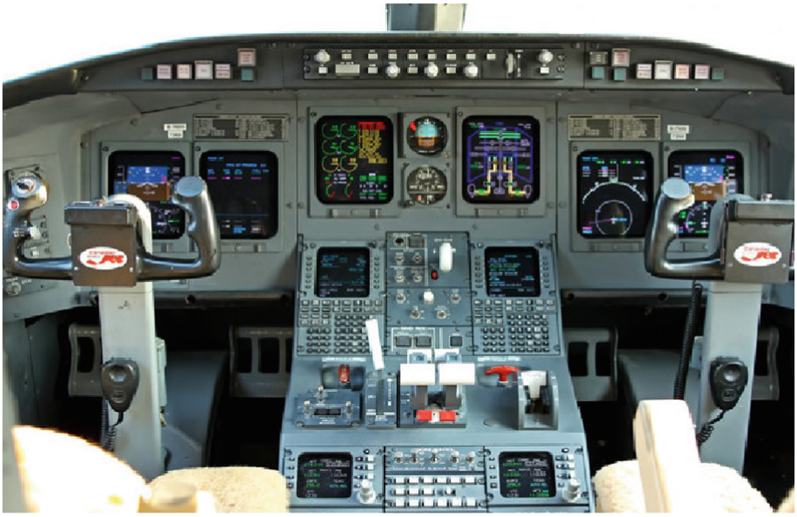


Fig. 1. Flight deck of CRJ-200

Besides the flight simulator and the aircraft, a head-mounted eye tracker (Tobbi Glass, Sweden), which sample rate was 30 Hz, was used to determine the eye blinks rate of the subjects during the experiment, shown as in Fig. 2.

Fig. 2. Tobbi Glass

2.3 Procedure

In order to compare the workload measurements in flight simulator and flight test, three flight scenarios were designed, including Standard Instrument Departure (SID), Manual Departure (MD), and Standard Instrument Approach (SIA). Each of the flight scenarios were carried out in flight simulator and flight test respectively by each flight crew. The configurations and operating procedures for the flight scenarios were same in flight simulator and flight test as following.

1. Standard Instrument Departure

The flight scenario was conducted in Chengdu Shangliu International Airport. The task was started from pressing "TOGA (Takeoff/Go-around)" button by pilots. Then, the subjects pushed the throttle and kept accelerating. When the aircraft reaching the speed of VR, the subjects needed to rotate and maintained a 3 degree climbing approximately. When the aircraft reaching the altitude of 1500 feet, the subjects were required to connect the autopilot system, and keep supervising the essential flight parameters until climbing to 10000 feet.

2. Manual Departure

The flight scenario was conducted in Chengdu Shangliu International Airport. The task was started from pressing "TOGA" button by pilots. Then, the subjects pushed the throttle and kept accelerating. When the aircraft reaching the speed of VR, the subjects needed to rotate and maintained a 3 degree climbing approximately. Moreover, when supervising the positive rising rate on Primary Flight Display, the subjects were required to retract the landing gear and keep climbing to 10000 feet by hand.

3. Standard Instrument Approach

The flight scenario was conducted in Chengdu Shangliu International Airport. The task was started in 40 nautical miles away from descending point. After slowing down to 145 knots, and descending to 1500 feet, the aircraft was in landing pattern. The subjects executed a CAT I standard instrument approach procedure and landed on the runway.

The simulation experiment was conducted prior to the flight test. At first time, the subjects performed a standard instrument departure and a standard instrument approach. At the second time, they performed a manual departure and a standard instrument approach. After each task, every subject was asked to fulfill the NASA-TLX scale. In flight test, the procedures were same as in flight simulator.

2.4 Statistical Analysis

SPSS 17.0 for Windows was used to process the experiment data, and ANOVA analysis, and correlation analysis were implemented in this study. When $P < 0.05$, the results were considered statistically significant.

3 Results

3.1 NASA-TLX Scales

Considering the results of NASA-TLX scales, the three flight scenarios showed the significant differences in the simulator experiment (F(2,12) = 3.01, p = 0.040). Among them, Standard Instrument Approach (SIA) had the maximum average NASA-TLX scores (Mean = 27.92, SD = 9.54), Standard Instrument Departure (SID) was minimum (Mean = 19.85, SD = 5.08), Manual Departure (MD) was in the middle (Mean = 22.58, SD = 7.32). Similarly, in flight test, standard instrument approach had the highest NASA-TLX scores (Mean = 33.42, SD = 10.24), manual departure was medium (Mean = 28.75, SD = 7.06), and standard instrument departure was minimum (Mean = 25.42, SD = 9.00). However, the differences of three flight scenarios in flight test were insignificant (F(2,12) = 3.01, p = 0.063). Furthermore, the difference between simulator experiment and flight test were significant in standard instrument departure (t = 2.43, p = 0.024) and in manual departure (t = 2.10, p = 0.047). Nevertheless, in standard instrument approach, the difference was insignificant (t = 1.36, p = 0.187). Otherwise, NASA-TLX scales showed a moderate correlation between simulator and flight test (R = 0.524, p = 0.001), as was depicted in Fig. 3.

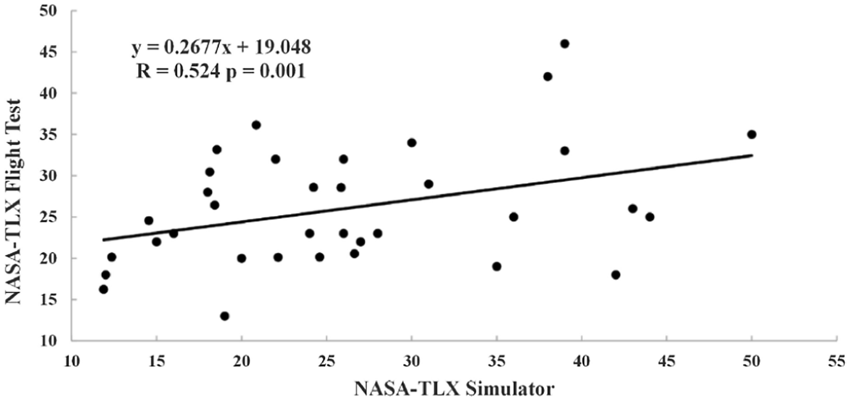


Fig. 3. The linear regression results of the NASA-TLX scales in simulator and in flight test

3.2 Eye Blinks Rate

Considering the results of eye blinks rate, as shown in Fig. 4, only in the simulator experiment (F(2,12) = 4.711, p = 0.016), the differences of the three flight scenarios was significant, and in the flight test (F(2,12) = 0.003, p = 0.997), the differences was insignificant. In the simulator experiment, standard instrument departure had the maximum average eye blinks rate (Mean = 14.08, SD = 3.63), standard instrument approach was minimum (Mean = 9.83, SD = 2.72), and manual departure was medium (Mean = 11.58, SD = 3.78). However, in the flight test, the discrepancy is slight. Furthermore, comparing the difference between simulator experiment and flight test for each flight scenarios respectively, only standard instrument departure was significant (t = 3.331, p < 0.01), and both manual departure (t = 1.457, p = 0.159) and standard instrument approach (t = 0.213, p = 0.834) were insignificant. Besides, eye blinks rate expressed a more weak correlation between simulator and flight test (R = 0.242, p = 0.155).

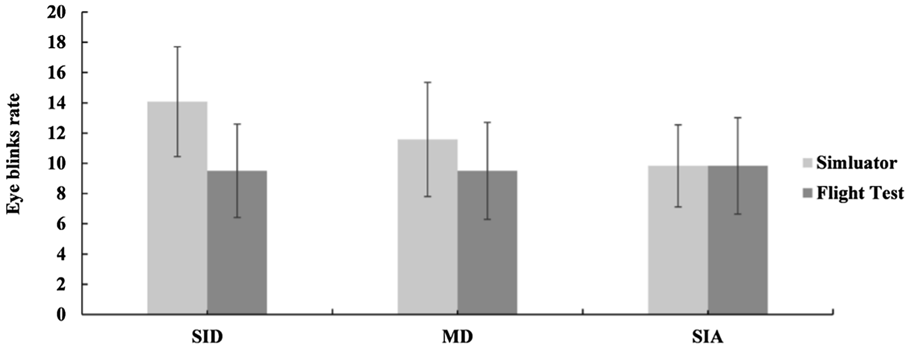


Fig. 4. The results of Eye blinks rate for the three flight scenarios, which were standard instrument departure (SID), manual departure (MD) and standard instrument approach (SIA), in the simulator and in the flight test. The error bars stand for the difference of eye blinks rate of the subjects either in simulator of in flight test.

4 Discussion

Flight test is the most direct means of compliance in aircraft human factors airworthiness certification. However, it is not the preferred means due to the following three reasons. Firstly, it might not be appropriate to test an abnormal situation for safety consideration [18]. Secondly, a flight environment is normally difficult to manipulate the operational environment which might be required to apply the scenario-based approach. Last but not least, human factors scenarios performed in flight test could not be easy to duplicate due to the lack of controllability of the operation context [19]. Therefore, simulator test might be more appropriate than flight test, especially in high risk flight scenarios, and both of them should be examined from the standpoint of human workload to shown compliance with airworthiness requirement.

However, the classic workload measurements have their own limitations. Subjective rating scale measures are sometimes uncertain on the repeatability and validity, and data manipulations are often questioned as being inappropriate [20]. Moreover, subjective feeling of workload was essentially dependent on the time stress involved in performing the task for time-stressed tasks only [21]. For task performance method, because of the compensatory effect of increased effort, it is clear t not sufficient to assess the state of the operator [22], and some other factors, such as strategy, affect performance and workload differently [23]. Psychophysiological measures are influenced by ambient environment and task duration [24]. In real flight, most of pilots are preferred to wear a sunglass to prevent direct sunlight. Moreover, some studies assumed that eye movement activity parameters only can provide a sensitive measure of visual workload [25]. Therefore, it is necessary to select the desirable workload measurements according to the specific characteristics of simulator test and flight test.

In the simulator experiment, NASA-TLX is a multidimensional rating scale that assesses a subject's subjective workload on six 100-point scales related to a different aspect of workload: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration [26]. It is more precise and comprehensive in workload evaluation. Besides, Eye measures were sensitive to intermediate levels of mental effort as well [27], and would also produce reliable near-real-time indicators of workload in flight simulator [28].

In this study, two types of workload measurements were compared, including subjective methods: NASA-TLX, and psychophysiological measures: eye blinks rate both in flight simulator and in flight test in three flight scenarios. The results demonstrated that NASA-TLX, eye blinks rate were credible in flight simulator. Nevertheless, in these three flight scenarios, neither of them produced reliable indictors in flight test. In further study, there are two more aspect would be carried out. Firstly, more measures would be implemented in both simulator and flight test environment, for instance subjective measurements including Bedford methods and Modified Cooper-Harper, and psychophysiological approaches like ECG and EEG. Secondly, in order to ensure the safety of flight, only normal flight scenarios were selected in this study. Therefore, under safe condition, more scenarios should be included, especially some abnormal conditions, such as, crosswind handling, one engine failure.

Acknowledgments. This research work was supported by Professor Fu, Dr. Lu and Dr. Wang in Man-Machine Environment lab, School of Electrical Engineering and Electronic Information, Shanghai Jiao Tong University.

References

1. De Florio, F.: Airworthiness: An Introduction to Aircraft Certification and Operations. Butterworth-Heinemann, Oxford (2016)
2. EASA: Certification Specifications for Large Aeroplanes CS-25 (2009)
3. Wiegmann, D.A., Shappell, S.A.: Human error analysis of commercial aviation accidents: application of the Human Factors Analysis and Classification System (HFACS). Aviat. Space Environm. Med. **72**(11), 1006–1016 (2001)

4. Farmer, E., Brownson, A.: Review of workload measurement, analysis and interpretation methods, vol. 33. European Organisation for the Safety of Air Navigation (2003)
5. Stone, G., Gulick, R., Gabriel, R.: Use of task timeline analysis to assess crew workload, DTIC Document (1987)
6. O'Donnell, R., Eggemeier, F.T.: Workload assessment methodology. Meas. Tech. **42**, 5 (1986)
7. Cain, B.: A review of the mental workload literature, DTIC Document (2007)
8. Ashcraft, M.H., Kirk, E.P.: The relationships among working memory, math anxiety, and performance. J. Exp. Psychol.: Gen. **130**(2), 224 (2001)
9. Bliss, J.P., Dunn, M.C.: Behavioural implications of alarm mistrust as a function of task workload. Ergonomics **43**(9), 1283–1300 (2000)
10. Wickens, C.D.: Multiple resources and mental workload. Hum. Factors: J. Hum. Factors Ergon. Soc. **50**(3), 449–455 (2008)
11. Wester, A., et al.: Event-related potentials and secondary task performance during simulated driving. Accid. Anal. Prev. **40**(1), 1–7 (2008)
12. Lansdown, T.C., Brook-Carter, N., Kersloot, T.: Primary task disruption from multiple in-vehicle systems. ITS J.-Intell. Transp. Syst. J. **7**(2), 151–168 (2002)
13. Schnell, T., et al.: Improved flight technical performance in flight decks equipped with synthetic vision information system displays. Int. J. Aviat. Psychol. **14**(1), 79–102 (2004)
14. Roper, D., et al.: Integrating CFD and piloted simulation to quantify ship-helicopter operating limits. Aeronaut. J. **110**(1109), 419–428 (2006)
15. De Waard, D.: The Measurement of Drivers' Mental Workload. Traffic Research Center, Groningen University, Netherlands (1996)
16. Hoover, A., et al.: Real-time detection of workload changes using heart rate variability. Biomed. Sig. Process. Control **7**(4), 333–341 (2012)
17. Tsai, Y.-F., et al.: Task performance and eye activity: predicting behavior relating to cognitive workload. Aviat. Space Environ. Med. **78**(5), B176–B185 (2007)
18. Schutte, P.C., Trujillo, A.C.: Flight crew task management in non-normal situations. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 244–248. SAGE, Los Angeles (1996)
19. Perkins, C.D.: Stability and Control: Flight Testing, vol. 2. Elsevier, Amsterdam (2014)
20. Annett, J.: Subjective rating scales: science or art? Ergonomics **45**(14), 966–987 (2002)
21. Meshkati, N., et al.: Techniques in mental workload assessment (1995)
22. Wilson, A.G.F.: Operator functional state assessment for adaptive automation implementation. In: Proceedings of SPIE - The International Society for Optical Engineering, vol. 5797, pp. 100–104 (2005)
23. Wickens, C.D., Huey, B.M.: Workload Transition: Implications for Individual and Team Performance. National Academies Press, Washington, DC (1993)
24. Allanson, J., Fairclough, S.H.: A research agenda for physiological computing. Interact. Comput. **16**(5), 857–878 (2004)
25. Wilson, G., Fisher, F.: The use of cardiac and eye blink measures to determine flight segment in F4 crews. Aviat. Space, Environ. Med. **62**(10), 959–962 (1991)
26. Hart, S.G.: NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Sage Publications (2006)
27. De Rivecourt, M., et al.: Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. Ergonomics **51**(9), 1295–1319 (2008)
28. Van Orden, K.F., et al.: Eye activity correlates of workload during a visuospatial memory task. Hum. Factors **43**(1), 111–121 (2001)

Group Collaboration and Decision Making

Spatial Ability in Military Human-Robot Interaction: A State-of-the-Art Assessment

Maartje Hidalgo(✉), Lauren Reinerman-Jones, and Daniel Barber

Institute for Simulation and Training, University of Central Florida,
32806 Orlando, FL, USA
{mhidalgo, lreinerm, dbarber}@ist.ucf.edu

Abstract. In the dismounted military field, robots and unmanned vehicles are increasingly used as force multipliers and teammates. As such, a fluent human-robot interaction (HRI) becomes vital and is stimulated by fitting the robot and its interface to the human teammate's capabilities. This is where individual differences of the human needs to be considered, such as those found in spatial ability. In HRI, information presented to the human teammate requires mental manipulation and interpretation to inform subsequent human actions, which relies on spatial ability. In order to generalize findings to the armed forces and to inform future design requirements, factors pertaining to construct operationalization, measurement, and task type need to be examined. The aim of the present literature review is to investigate spatial ability findings in military HRI.

In this review, metadata over the past decade are synthesized in light of a formal factor analysis of spatial ability [8]. The results show that the operationalizations of spatial ability are alarmingly divergent in the research field of military/UxV HRI. The relationship between spatial ability and task performance in our findings is complicated by a lack of standardized assessments of spatial ability and a small sample size, which is an indication of the current state of affairs. However, there is a conservative indication of a relationship between aspects of spatial ability and primary military reconnaissance tasks. As an effort to inform future studies, this literature review concludes with recommendations for military-affiliated research and development, to enhance the measurement, validation, and generalization of findings of an individual factor that has the potential to benefit a fluent HRI and the transition of robots from tools to teammates.

Keywords: Human-robot interaction · Applied cognitive psychology · Human factors/system integration

1 Introduction

Research in human-robot interaction (HRI) shows the recent transition for robots, from tools to teammates [1]. Aside from dismounted military members working with physical robots, a development is growing with the implementation of unmanned aerial (UA) and ground vehicles (UGVs) [2–5]. The interaction with robots is pursued through augmentation of interfaces. Interpretation of the information

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 363–380, 2019.
https://doi.org/10.1007/978-3-030-22507-0_29

displayed through such interfaces often requires mental manipulation and interpretation by the human teammate or operator. This changes the work parameters of operators significantly and leads to the importance of considering individual differences in the ability to cognitively process input from the robot/UxV adequately [6]. This process occurs through mental rotation and visualization, both aspects of spatial ability [7]. However, the exact role is still largely unknown. What complicates research of the role of spatial ability, is that the construct is still unclear, despite decades of research.

1.1 Spatial Ability

Factor analyses have shown that spatial ability is a complex construct. In this section, we highlight two major efforts to discuss facets of spatial ability, one by Lohman, Pellegrino, Alderton, and Regian [7] and one by Carroll [8]. Table 1 summarizes and contrasts their factors and definitions of spatial ability.

Lohman and colleagues [7] identified ten factors of spatial ability, and Lohman summarizes this human aptitude as the "ability to generate, retain, retrieve, and transform well-structured visual images" [9] (p. 3). By contrast, Carroll's [8] definition of spatial ability emphasizes a visual component: "spatial and other visual perceptual abilities have to do with individuals' abilities in searching the visual field, apprehending the forms, shapes, and positions of objects as visually perceived, forming mental representations of those forms, shapes, and positions, and manipulating such representations mentally" (p. 304). His factor analysis of spatial ability classified five major factors (Table 1). The most complex factor is *Visualization* with several task types that load onto it, including block rotation tasks, assembly tasks, paper folding tasks, perspective-taking tests, and tasks that require participants to determine how an object should be formed into or broken down from a three-dimensional image [8, 10]. *Spatial Relations,* according to Carroll, constitutes Lohman et al.'s factor *Spatial Orientation*, with an emphasis on the speed in which the individual is capable of mentally manipulating simple visual patterns, such as determining whether one pattern is a rotated version of another image.

Recent research still scrutinizes these factors and newer measures of spatial ability have been generated [11, 12], indicating that congruence over the constituents of spatial ability has not been reached. This complicates the evaluation of the effect of spatial ability in human-robot interaction as well. However, as Carroll's [8] analysis is the most recent factor analysis of spatial ability to date, this framework will be used in the remainder of this paper.

Table 1. Operationalizations and factors of spatial ability [7, 8].

Lohman et al. [7]		Carrol [8]	
Factor	Operationalization	Factor	Operationalization
Visualization	Folding, mentally transforming, rotating objects	*Visualization*	Perceiving, mentally manipulating, matching, and mentally rotating objects, as well as breaking forms down or forming into whole form
Spatial orientation	Viewing object/scene from different perspective	*Spatial relations*	Mentally manipulating and matching patterns or objects
Flexibility of closure	Breaking down a whole form to create a new form	*Closure flexibility*	Recognizing a known, non-hidden, visual pattern
Closure speed	Speed of identifying that a form is incomplete or misrepresented	*Closure speed*	Speed of recognizing a concealed, unknown visual pattern
Perceptual speed	Speed of matching figures	*Perceptual speed*	Speed of recognizing a known, non-hidden, visual pattern
Spatial scanning	Maze-tracing, which also loads on spatial planning and perceptual speed factors		
Serial integration	Integrating visual stimuli based on short-term memory		
Visual memory	Recognizing an image (from memory) that was previously presented		
Kinesthetic	Self-orientation in space, especially left-right discriminations		

1.2 Purpose

The purpose of this paper is to review how spatial ability has been operationalized and measured in research of human-robot interaction (HRI) in the military field, including unmanned vehicles. This is a state-of-the-art assessment, focused on construct operationalization, measurement, and applied task types. The metadata will be synthesized to determine the degree of vergence in the assessment of spatial ability in military HRI research, and how this relates to task performance. The goal is to assess the generalizability of the current modus operandi and to generate recommendations where appropriate.

2 Method

2.1 Information Sources and Inclusion Criteria

The databases Compendex/EngineeringVillage, PsycInfo, and ProQuest were accessed to locate articles that were published in the last 10 years. The following keywords were all required for a study to be included in the qualitative analysis: (a) human-robot interaction (HRI) OR human-agent teaming; AND (b) military OR unmanned vehicle (UxV); AND (c) spatial ability.

2.2 Exclusion Criteria

Studies that included a robot but were not in the realm of the military or unmanned vehicles were excluded, such as assistance and surgical robots. Additionally, only studies that presented primary data from journal or conference papers were included. Papers primarily focused on predictive modeling were excluded as well, as the purpose of the present paper is to relate spatial ability, and its measures, to task performance. The remaining analyzable sample size was N = 23.

2.3 Data Extraction

The following data were extracted for analysis: labeled construct (i.e. the name assigned to the construct by the authors), spatial ability test, task type, and task performance outcome. Any effects on workload, situation awareness, or other outcome measures, were not considered. All data were imported into an Excel worksheet to categorize and qualitatively analyze the data.

3 Results

To evaluate the relationship between spatial ability and performance in military/UxV human-robot interaction, the metrics and constructs of spatial ability were qualitatively analyzed. Next, the tasks in which spatial ability was evaluated were categorized and the metadata were synthesized and related to performance outcome.

3.1 Spatial Ability Metrics

The reviewed publications utilized ten different tests of spatial ability. Figure 1 displays the frequency of the different tests, as well as an overview of the labels authors assigned to the construct associated with the test. The three most used tests were the Cube Comparison Test [13] (CCT), Spatial Orientation Test [14] (SOT), and Mental Rotation Test [15] (MRT). The labels that were most often used by researchers for the tests were “spatial ability” (SpA), “spatial orientation” (SpO), and “spatial visualization” (SpV). The next section will provide a more in-depth analysis the tests and each of these three labels.

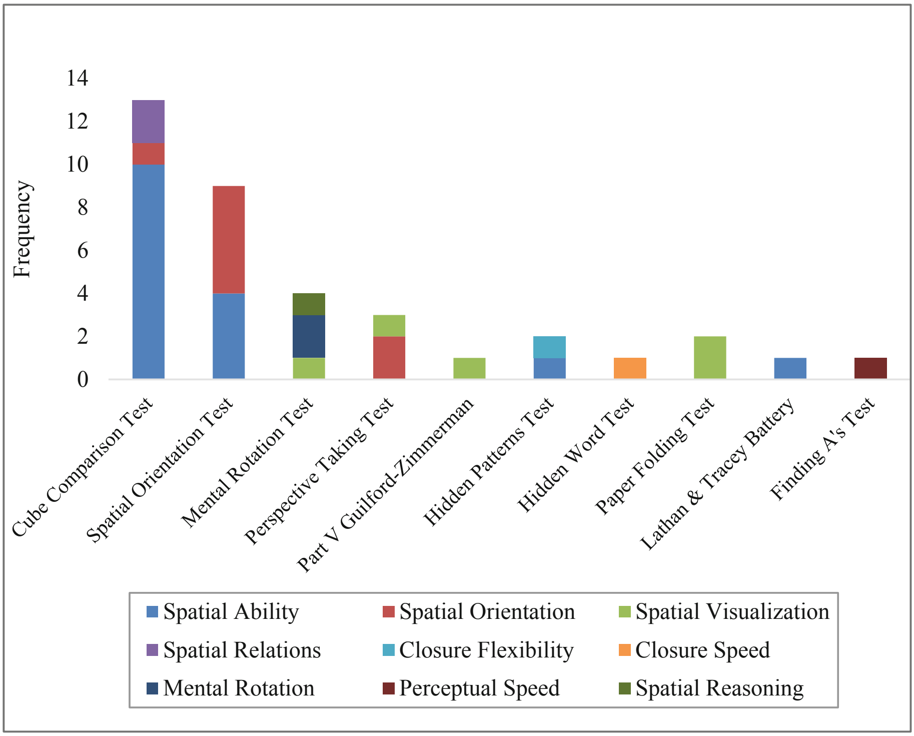


Fig. 1. Tests of spatial ability in across included studies. Legend shows the labels authors assigned to the constructs measured by the tests.

3.2 Spatial Ability Labels

Table 2 provides an overall summary of the ten spatial ability tests, the associated researcher-assigned label which is contrasted against the construct as categorized by Carroll's [8] factor analysis, and a frequency of use of the test and label in the current sample. In this section, the three most used labels by researchers (SpA, SpO, SpV) will be discussed further.

Table 2. Overview of tests of spatial ability, label associated with the test, the construct associated with the test based on Carroll [8], and frequency of use in the current sample. Tests that were developed after Carroll's analysis are coded as N/A for the construct.

Test	Label	Construct based on Carroll [8]	Frequency
Cube Comparison Test [13] (CCT)	Spatial ability	*Visualization*	10
	Spatial orientation	*Spatial relations*	1
	Spatial relations		2

(*continued*)

Table 2. (*continued*)

Test	Label	Construct based on Carroll [8]	Frequency
Spatial Orientation Test [14] (SOT)	Spatial ability	N/A	4
	Spatial orientation		5
Perspective Taking Test [11, 12] (PTT)	Spatial orientation	N/A	2
	Spatial visualization		1
Part V Guilford-Zimmerman [16] (V-GZ)	Spatial visualization	*Visualization*	1
Mental Rotations Test [15] (MRT)	Spatial visualization	*Visualization*	2
	Mental rotation		2
	Spatial reasoning		1
Paper Folding Test [13] (PFT)	Spatial visualization	*Visualization*	3
Hidden Patterns Test [13] (HPT)	Spatial ability	*Closure flexibility*	1
	Closure flexibility		1
Hidden Words Test [13] (HWT)	Spatial ability	*Closure speed*	1
Lathan & Tracey Battery [17] (LTB)	Spatial ability	N/A	1
Finding A's Test [13] (FAs)	Perceptual speed	*Perceptual speed*	1

Spatial Ability (SpA). SpA was measured 13 times in the current sample (N = 23) with four different tests: CCT, SOT, Hidden Patterns Test [13] (HPT), and Lathan and Tracey's test battery [17] (LTB). When looking at the term spatial ability, ideally more than one test would be used to measure it, considering its multifaceted nature, such as in LTB. However, in the current sample, 40% of the studies implement one test to assess (a form of) spatial ability. The tests were further examined to compare them to Carroll's [8] factor analysis.

In the CCT, individuals must evaluate a rotated cube and determine whether or not it is similar to the target cube. Carroll [8] determined that this test loads on the factor *Visualization* and *Spatial Relations* (Table 1).

The SOT is based on Gugerty and Brooks' [14] cardinal direction task, wherein participants are presented with a north-up map that shows the location and (varied) heading of their aircraft and a ground target. Additionally, they see a forward view from the perspective of the aircraft, displaying the ground target, a building, and four parking lots around the building. Participants are required to indicate in which cardinal direction the parking lot with vehicles is (only one): north, east, south, or west of the building. This test was developed after Carroll's [8] factor analysis; thus formal factor analysis was not applied. However, the perspective-taking properties suggest the test may load on the Carroll's factor *Visualization* and potentially *Spatial Relations.*

The HPT is a task wherein participants are to determine whether or not a specific orientation occurs for the target item. This test loads on the factor *Closure Flexibility* [8].

The LTB refers to the CTY Spatial Test Battery [18], which includes a complex figure test, a spatial memory test, a block rotation test, and a cube perspective test. This battery was not evaluated by Carroll [8] but incorporates several facets of *Visualization*.

Spatial Orientation (SpO). SpO was measured eight times with four different tests, from highest to lowest count: SOT, Perspective Taking Test [11, 12] (PTT), and CCT. As the SOT and CCT have been discussed, only the PTT will be discussed in this section.

The PTT is a more recent test that was developed based on evidence for a dissociation between mental rotation and perspective-taking spatial abilities [11]. In this task, individuals are asked to imagine different perspectives or orientations in space, imaginatively standing at one object while facing another object, and then pointing to where a third object would be. This test was developed after Carroll's [8] factor analysis. As a perspective-taking task, this test would most likely be categorized as a factor of *Visualization* and *Spatial Relations*, although factor analysis needs to confirm this.

Spatial Visualization (SpV). In the current sample SpV was measured seven times with five different tests: Paper Folding Test [13] (PFT), Part V of the Guilford-Zimmermann Aptitude Test [16] (V-GZ), Mental Rotations Test [15] (MRT), and the PTT, which has been discussed in the previous section.

The PFT shows successive drawings of a folded paper. The last image shows where a hole is punched. The individual then needs to select one out of five options to indicate where the hole would be if the paper is unfolded. This loads on the factor *Visualization* [8].

The V-GZ asks individuals to look at perspective scenes as if they are looking over the prow of a boat and to indicate how the boat moved between two views. Carroll [8] categorized this task as a *Visualization* task. The task was not analyzed for the *Spatial Relations* factor, as Carroll followed the lead of others such as Ekstrom and colleagues [13].

The MRT presents an image consisting of multiple stringed cubes, in different rotational angles. The individual needs to indicate which two of four images are rotated images of the target picture. As a block rotation task, his test loads on the factor *Visualization* [8].

3.3 Task Analysis

When evaluating the effect of spatial ability on HRI task performance, task type needs to be considered. Task type defines the specific capabilities required from the human teammate. Therefore, tasks were classified into categories, as shown in Table 3, matched with the spatial ability test used, Carroll's [8] associated construct with the test, and a frequency within the current sample of studies. As expected from military HRI research, tasks are indeed representative of military reconnaissance missions.

Table 3. Categorization of task type with the applied test of spatial ability, associated construct based on Carroll's [8] factor analysis, and a frequency of use in the present sample. Tests that were developed after Carroll's analysis are coded as N/A for the construct.

Task type	Test	Construct based on Carroll [8]	Frequency
Robot control/teleoperation	CCT	*Visualization spatial relations*	11
	PTT	N/A	2
	SOT	N/A	9
	MRT	*Visualization*	2
	LTB	N/A	1
	PFT	*Visualization*	1
Target detection	CCT	*Visualization spatial relations*	9
	SOT	N/A	10
	V-GZ	*Visualization*	1
	MRT	*Visualization*	2
	HPT	*Closure flexibility*	3
	HWT	*Closure speed*	1
	FAs	*Perceptual speed*	1
Communication	CCT	*Visualization spatial relations*	5
	SOT	N/A	5
	HPT	*Closure flexibility*	3
Route planning	CCT	*Visualization spatial relations*	2
	PFT	N/A	1
	SOT	N/A	2
	V-GZ	*Visualization*	1
	MRT	*Visualization*	1
	HWT	*Closure speed*	1
	FAs	*Perceptual speed*	1
Target encapsulation	CCT	*Visualization spatial relations*	1
	SOT	N/A	1
Tacton classification	CCT	*Visualization spatial relations*	2
Respond system warnings	PFT	N/A	1
Sensor control	PFT	N/A	1

Robot control/teleoperation and target detection tasks were most frequently used, presumably as they are primary tasks in military reconnaissance missions with human-robot teams. The ability to perform robot control/teleoperation tasks has been evaluated with the spatial factors *Visualization* and *Spatial Relations*, which represent the ability to perceive and to mentally manipulate/rotate images to make a decision or judgment call for the next action, as well as the speed at which this is performed (Table 1).

Target detection ability in military HRI has been primarily associated with the spatial factor *Visualization,* followed by *Spatial Relations, Closure Flexibility*, *Closure Speed* and *Perceptual Speed*. Target detection indeed is a complex action sequence that involves perception of the environment, potentially mentally rotation depending on the field of view, as well as searching for and discriminating targets, and then marking them in the simulation. This sequence action is likely to hinge on speed of execution as well.

Secondary tasks of the simulated HRI missions in this sample include communication, route planning, target encapsulation, tacton classification when using tactile means of communicating, responding to warning systems, and controlling sensors. These tasks were associated with the spatial ability factors *Visualization, Spatial Relations, Perceptual Speed,* and *Closure Flexibility.*

3.4 Performance Outcome

Table 4 summarizes the key findings for task performance in relation to spatial ability, with a representation of the metrics used to measure (aspects of) spatial ability.

Table 4. Integration of relevant key findings for spatial ability according to task type.

Author	Year	Test SpA	Task type	Results
Chen and Barnes [19]	2008	CCT HPT SOT	Robot control/tele-operation Target detection Communication	Participants with higher spatial ability performed better on target detection tasks and teleoperation
Chen, Durlach, Sloan and Bowens [20]	2008	CCT	Robot control/tele-operation Target detection	Participants with higher spatial ability performed better on target detection tasks
Kumar and Sekmen [21]	2008	MRT	Robot control/tele-operation	High spatial reasoning ability correlated to higher robot control/teleoperation performance
Chen [22]	2009	CCT SOT	Robot control/tele-operation Target detection Communication	Participants with higher spatial ability performed better on target detection tasks. Communication performance not correlated to spatial ability
Chen and Joyner [23]	2009	CCT SOT	Robot control/tele-operation Target detection	SOT scores were the most accurate predictor of target detection performance, with a higher score correlated with higher task

(*continued*)

Table 4. (*continued*)

Author	Year	Test SpA	Task type	Results
			Communication	performance. SOT and CCT not correlated with robot control/teleoperation
Chen and Terrence [24]	2009	CCT SOT	Robot control/tele-operation Target detection Communication	Participants with higher spatial ability performed better on target detection tasks
Long, Gomer, Moore and Pagano [25]	2009	CCT PFT	Robot control/tele-operation	Only spatial relations ability positively correlated with teleoperation performance. Both spatial visualization and spatial relations test scores positively correlated with direct robot control
Chen [26]	2010	SOT	Robot control/tele-operation Target detection	Participants with higher spatial ability performed better on target detection tasks and teleoperation when they used a large UAV rather than smaller ones, especially when combined with a fixed rather than rotating view
Fincannon, Ososky, Jentsch, Keebler and Phillips [27]	2010	V-GZ	Target detection Route planning	Participants with higher spatial ability performed better on target detection tasks
Baber, Morin, Parkh, Cahillane and Houghton [28]	2011	PFT	Sensor control task Target classification task Respond to system warnings	Spatial orientation capabilities associated with time to respond to warnings when controlling multiple UGVs: Lower ability was associated with slower reaction time
Chen [29]	2011	CCT SOT	Robot control/tele-operation Target detection	Participants with higher spatial ability performed better on target detection tasks and teleoperation, the latter only when there is no aided target recognition
Chien, Wang and Lewis [30]	2011	PFT SOT	Robot control/tele-operation Route planning	No reported results for the described tasks. Spatial ability was correlated to other metrics not evaluated in the present effort (marking map for location of victims)

(*continued*)

Table 4. (*continued*)

Author	Year	Test SpA	Task type	Results
Gomer and Pagano [31]	2011	CCT PFT	Robot control/tele-operation	Higher spatial ability, as a composite score, was correlated with better robot control/teleoperation performance. Spatial visualization may be a unique contributor to robot control/teleoperation
Long, Gomer, Wong and Pagano [32]	2011	CCT PFT	Robot control/tele-operation	Participants with higher spatial ability performed better on robot control. Higher spatial relations ability was positively correlated with teleoperation
Blitch, Bauder, Gutzwiller, and Clegg [33]	2012	LTB	Robot control/tele-operation	Higher spatial ability correlated to faster task completion time
Chen and Barnes [34]	2012a	CCT SOT	Robot control/tele-operation Target detection Route planning	Participants with higher spatial ability performed better on target detection tasks. Participants with higher spatial ability performed better on route planning tasks
Chen and Barnes [35]	2012b	CCT SOT	Target encapsulation Secondary target detection	Spatial ability not correlated with target encapsulation performance. Participants with higher spatial ability performed better on secondary target detection tasks
Fincannon, Jentsch, Sellers and Keebler [36]	2012	MRT HWT FAs	Target detection Route planning	Higher visualization and closure speed abilities correlated with better target detection performance
Fincannon, Keebler, Jentsch and Curtis [6]	2013	MRT	Target detection	Participants with higher spatial ability performed better on target detection tasks
Barber, Reinerman-Jones and Matthews [37]	2015	CCT	Tacton classification	Spatial ability was correlated with classification accuracy for dynamic tactons and static tactons, but not with directional tactons Spatial ability was negatively correlated with reaction time for a single word and tacton phrases. Spatial ability correlated positively with classification accuracy for single words, but not for tacton phrases

(*continued*)

Table 4. (*continued*)

Author	Year	Test SpA	Task type	Results
Abich and Barber [38]	2017	CCT	Robot control/operation Route planning	No reported results for the described tasks. Spatial ability was correlated to other metrics not evaluated in the present effort
Reinerman-Jones, Barber, Szalma and Hancock [39]	2017	CCT	Tacton classification In Exp. 2: Robot control/tele-operation	Spatial ability not correlated with tacton classification performance
Wright, Chen and Barnes [40]	2018	SOT	Robot control/tele-operation Target detection Route planning	Participants with higher spatial ability showed higher sensitivity to target presence (target detection performance). Spatial ability not related to response bias

Table 4 indicates that, in general, primary military reconnaissance tasks are solely evaluated in relation to spatial ability, that is target detection and robot control/teleoperation performance. Secondary tasks such as communication or route planning are seldom evaluated in relation to spatial ability and generally did not show a significant relationship. The relatively small sample size and lack of congruence in utilized tests complicates a clear synthesis of the results.

Based on the findings from the studies in this sample, target detection performance seems to benefit from higher spatial ability. Several studies found a positive relation with capabilities represented by the factors *Visualization* and *Spatial Relations*. These findings imply that target detection relies, at least in part, on the ability to perceive, mentally manipulate and rotate, and discriminate or match visual images in a dynamic environment. Perspective-taking ability, through scene perception or block rotation, may be important additional aspects of this aptitude.

Robot control/teleoperation performance generally appears to be fostered by identical factors, i.e., *Visualization* and *Spatial Relations*. However, some of the reviewed studies did not find a significant relationship between the score on the spatial ability test and task performance [23], or only when an aid for the target detection performance task was not available [29]. Hence, there may be a weaker relationship between spatial ability and robot control/teleoperation performance than with target detection performance. Another possibility is that a significant relationship was not found due to the perspective used in the task. Indeed, a number of studies made a distinction between direct robot control and remote robot control through teleoperation, and found that teleoperation performance was specifically related to the CCT, while a composite of the CCT and PFT was related to direct robot control performance [25, 31, 32]. This finding indicates that the perspective used in the task may be a moderating factor in the relationship between spatial ability and robot control/teleoperation.

4 Discussion

The purpose of this paper is to review how spatial ability has been researched in military human-robot interaction, including unmanned vehicles (military/UxV HRI). Spatial ability is a multi-faceted construct [7, 8]. In this review, we focused on the factors of spatial ability as defined in the factor analysis by Carroll [8], which focuses on five factors of visualization, as part of spatial ability and as important aspects of interfacing [41]. Spatial ability, according to Carroll, emphasizes “individuals’ abilities in searching the visual field, apprehending the forms, shapes, and positions of objects as visually perceived, forming mental representations of those forms, shapes and positions, and manipulating such representations mentally” (p. 304), and is factor analytically decomposed in the factors *Visualization, Spatial Relations, Closure Speed, Closure Flexibility*, and *Perceptual Speed*. The goal of this literature review is to provide a state-of-the-art assessment of spatial ability to assess the generalizability of the current modus operandi and to generate recommendations for the highly specialized field of military/UxV HRI.

The first interesting finding is the relative scarcity of peer-reviewed journal publications, that provide primary data, in the realm of military/UxV HRI that analyze the effect of spatial ability. In a search of the recent decade, 23 viable publications were identified for qualitative analysis. The labels that were most often assigned by researchers to the measures of spatial ability were “spatial ability,” “spatial orientation,” and “spatial visualization.”

The term “spatial ability” is used fairly loosely in this sample. It was the most common measured label and was assessed with four different tests. When “spatial ability” is measured, several tests are required to adequately reflect the multifaceted nature of the construct. In this sample, 60% combined three tests at most when measuring “spatial ability”. This indicates that a large portion used merely one test when claiming to measure “spatial ability”, which connotes a lack of construct validity.

Furthermore, there is a lack of convergence in the definition, i.e., researcher-assigned labels, to the tests. These labels are often not verified by formal factor analysis that would strengthen the validity. One example is the most common used measure in this sample, the Cube Comparison Test [13], which the majority of the researchers labeled as a metric of “spatial ability.” A few researchers more accurately defined this test as “spatial visualization” and “spatial relations”, conform Carroll’s [8] extensive factor analysis. Another frequently used test is the Spatial Orientation Test [14], which was published after Carroll’s factor analysis. Even though the name of the test suggests would be a factor of *Spatial Relations/Orientation*, the fact that this test requires participants to shift perspectives, implies the task may be a factor of *Visualization* instead [8, 10].

Next, task types were analyzed in relation to spatial ability. The work parameters of the human interfacing with a robot (or unmanned vehicles) are different from the human actually being in the field. Interfacing with robots relies on the ability to mentally rotate and visualize the information presented through the interface [8, 41]. Therefore, there is a need to analyze the relationship between spatial ability factors and task types, especially in relation to performance. Task type defines the specific capabilities required from the human teammate and therefore should inform which spatial ability tests to implement in the assessment.

The task types used in this sample are similar to military reconnaissance tasks as applied to interactions with robots or unmanned vehicles. The primary tasks are robot control/teleoperation and target detection, with secondary tasks as route planning, communication, responding to system warnings, and tacton classification. In general, tests of *Visualization* and *Spatial Relations* have been indexed in relation to robot control/teleoperation tasks. Target detection tasks were represented by several factors of spatial ability, including *Visualization*, *Spatial Relations*, *Closure Flexibility*, *Closure Speed*, and *Perceptual Speed*. Target detection tasks may be more complex in nature, therefore relying on different aspects of spatial ability. Other factors that were not reviewed may also play a role in the versatility of the tests used, such as frame of reference of the robot/UxV and interface [41]. Furthermore, although secondary tasks were listed, few studies evaluated spatial ability in relation to secondary task performance. The secondary tasks were perhaps employed to simulate the complexity of a dismounted military reconnaissance scenario, wherein primary task performance is more critical to the mission.

Lastly, the performance outcomes were synthesized and evaluated to find potential patterns. Considering the relatively small sample size and lack of congruence in utilized tests, a generalization of the results was complex and the findings should be interpreted with great caution. There is a conservative indication that performance on robot control/teleoperation and target detection tasks relies at least on some form of spatial ability. Target detection performance may benefit from the ability to perceive, mentally manipulate and rotate, and discriminate or match visual images in a dynamic environment. Perspective-taking ability, through scene perception or block rotation, may be a factor that plays a role in target detection as well. Furthermore, in the current sample, robot control/teleoperation performance does not consistently show a relationship with higher spatial ability. Specifically, there may be a distinction between the aspects of spatial ability that affect direct robot control versus remote robot teleoperation. Thus, robot control/teleoperation performance may depend in some ways on the perspective or frame of reference used in the task.

4.1 Future Research

Based on this literature review, the scientific community is recommended to focus on creating a congruent foundation of assessments in the critical and specialized field of military/unmanned vehicle HRI. We need a general awareness of the factors that form spatial ability, based on validated factor analyses, and how these factors relate to the performance type that is required of the human based on the task. By creating congruence and specificity in how factors of spatial ability are measured, findings can be replicated, which contributes to the foundation of theory in this fast-growing field. Replicated findings, with consistent measures, contributes to the reliability and validity of findings, which is vital when the findings are to be generalized to real-world military missions.

Additionally, the synthesis of the metadata shows at least an indication of a relationship between certain aspects of spatial ability and primary military reconnaissance task performance. Future research should attempt to understand this relationship on a deeper level, by evaluating the effects of spatial ability in such tasks with consistency and with tests that are reflective of factors that are relevant to the context. Replication and validation are key. Furthermore, with a larger database of replicated findings formal meta-analysis can be conducted, which yields statistically tested and generalizable results.

Finally, when looking at the metadata, it seems that researchers tend to have a preference for the specific measures of spatial ability that are implemented. To some extent, this may be a valid choice based on the task paradigm used. However, researchers need to remain cautious of a researcher bias when selecting assessment metrics.

4.2 Limitations

The present study is limited in its scope and size, as only studies that relate to military operations and include robots or unmanned vehicles were included. This limits the generalizability of the results, although specificity is maintained. A selection bias may have occurred due to the small, specialized field in which the literature review was conducted. Quality of the results was attempted to be maintained by only selecting peer-reviewed journals. Furthermore, no formal statistical analyses were performed; therefore the results should be interpreted with caution and only be taken as an indication.

5 Conclusion

To enhance a fluent interaction between human and robot teammates in the military field, future design requirements need to be informed by the human teammate's capabilities. Individual differences in spatial ability require special attention, as human-robot interfacing relies on the ability to mentally manipulate and interpret the communicated information to inform subsequent actions. The purpose of this literature review is to determine how spatial ability has been operationalized military/unmanned vehicle human-robot interaction, focusing on construct operationalization, measurement, and applied task types. Synthesis of metadata of the sample from the past decade shows that the operationalizations of spatial ability in this highly specialized field are divergent, with evidence of issues with construct validity. There is preliminary evidence for a relationship between aspects of spatial ability and primary military reconnaissance task performance. This relationship should be further investigated with consistent and validated measures of spatial ability that are selected based on the task type and demands posed on the human teammate, so that reliable generalizations can be made real-world military missions.

References

1. Phillips, E., Ososky, S., Grove, J., Jentsch, F.: From tools to teammates: toward the development of appropriate mental models for intelligent robots. Presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 55, pp. 1491–1495 (2011)
2. Blair, E.A., Rahill, K.M., Finomore, V., Satterfield, K., Shaw, T., Funke, G.: Best of both worlds: evaluation of multi-modal communication management suite. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 58, pp. 410–414 (2014)
3. Burnett, G.M., Wischgoll, T., Finomore, V., Calvo, A.: Multimodal mobile collaboration prototype used in a find, fix, and tag scenario. In: Uhler, D., Mehta, K., Wong, J.L. (eds.) MobiCASE 2012. LNICST, vol. 110, pp. 115–128. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36632-1_7
4. Calvo, A.A., Finomore, V.S., Burnett, G.M., McNitt, T.C.: Evaluation of a mobile application for multimodal land navigation. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 57, pp. 1997–2001 (2013)
5. Kim, S., Miller, M.E., Rusnock, C.F., Elshaw, J.J.: Spatialized audio improves call sign recognition during multi-aircraft control. Appl. Ergonomics **70**, 51–58 (2018)
6. Fincannon, T., Keebler, J.R., Jentsch, F., Curtis, M.: The influence of camouflage, obstruction, familiarity and spatial ability on target identification from an unmanned ground vehicle. Ergonomics **56**(5), 739–751 (2013)
7. Lohman, D.F., Pellegrino, J.W., Alderton, D.L., Regian, J.: Dimensions and components of individual differences in spatial abilities. In: Irvine, S.H., Newstead, S.E. (eds.) Intelligence and Cognition: Contemporary Frames of Reference. ASID, vol. 38, pp. 253–312. Springer, Dordrecht (1987). https://doi.org/10.1007/978-94-010-9437-5_6
8. Carroll, J.B.: Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press, New York (1993)
9. Lohman, D.F.: Spatial ability and g. In: Dennis, I., Tapsfield, P. (eds.) Human Abilities: Their Nature and Measurement, pp. 97–116. Lawrence Erlbaum Associates Inc, Mahwah (1996)
10. Fincannon, T.: Visuo-spatial abilities in remote perception: a meta-analysis of empirical work. Dissertation, University of Central Florida, Orlando, FL (2013)
11. Hegarty, M., Waller, D.: A dissociation between mental rotation and perspective-taking spatial abilities. Intelligence **32**(2), 175–191 (2004)
12. Kozhevnikov, M., Hegarty, M.: A dissociation between object manipulation spatial ability and spatial orientation ability. Memory Cogn. **29**(5), 745–756 (2001)
13. Ekstrom, R.B., Dermen, D., Harman, H.H.: Manual for Kit of Factor-Referenced Cognitive Tests, vol. 102. Educational Testing Service, Princeton (1976)
14. Gugerty, L., Brooks, J.: Reference-frame misalignment and cardinal direction judgments: group differences and strategies. J. Exp. Psychol.: Appl. **10**(2), 75 (2004)
15. Vandenberg, S.G., Kuse, A.R.: Mental rotations, a group test of three-dimensional spatial visualization. Percept. Mot. Skills **47**(2), 599–604 (1978)
16. Guilford, J.P.: The Guilford-Zimmerman aptitude survey. Pers. Guidance J. **35**(4), 219–223 (1956)
17. Lathan, C.E., Tracey, M.: The effects of operator spatial perception and sensory feedback on human-robot teleoperation performance. Presence: Teleoperators Virtual Environ. **11**(4), 368–377 (2002)
18. Eliot, J., Stumpf, H.: CTY Spatial Test Battery. Center for Talented Youth, Baltimore (1992)

19. * Chen, J.Y., Barnes, M.J.: Robotics operator performance in a military multi-tasking environment. Presented at the Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, pp. 279–286 (2008)
20. * Chen, J.Y., Durlach, P.J., Sloan, J.A., Bowens, L.D.: Human–robot interaction in the context of simulated route reconnaissance missions. Mil. Psychol. **20**(3), 135–149 (2008)
21. * Kumar, S., Sekmen, A.: Single robot - multiple human interaction via intelligent user interfaces. Knowl.-Based Syst. **21**(6), 458–465 (2008)
22. * Chen, J.Y.C.: Concurrent performance of military and robotics tasks and effects of cueing in a simulated multi-tasking environment. Presence Teleoperators Virtual Environ. **18**(1), 1–15 (2009)
23. * Chen, J.Y., Joyner, C.T.: Concurrent performance of gunner's and robotics operator's tasks in a multitasking environment. Mil. Psychol. **21**(1), 98–113 (2009)
24. * Chen, J.Y.C., Terrence, P.I.: Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. Ergonomics **52**(8), 907–920 (2009)
25. * Long, L.O., Gomer, J.A., Moore, K.S., Pagano, C.C.: Investigating the relationship between visual spatial abilities and robot operation during direct line of sight and teleoperation. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 53, pp. 1437–1441 (2009)
26. * Chen, J.Y.C.: UAV-guided navigation for ground robot tele-operation in a military reconnaissance environment. Ergonomics **53**(8), 940–950 (2010)
27. * Fincannon, T.D., Ososky, S., Jentsch, F., Keebler, J., Phillips, E.: Some good and bad with spatial ability in three person teams that operate multiple unmanned vehicles. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 54, pp. 1615–1619 (2010)
28. * Baber, C., Morin, C., Parekh, M., Cahillane, M., Houghton, R.J.: Multimodal control of sensors on multiple simulated unmanned vehicles. Ergonomics **54**(9), 792–805 (2011)
29. * Chen, J.Y.C.: Individual differences in human-robot interaction in a military multitasking environment. J. Cogn. Eng. Decis. Making **5**(1), 83–105 (2011)
30. * Chien, S.-Y., Wang, H., Lewis, M.: Effects of spatial ability on multi-robot control tasks. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 55, pp. 894–898 (2011)
31. * Gomer, J.A., Pagano, C.C.: Spatial perception and robot operation: should spatial abilities be considered when selecting robot operators? In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 55, pp. 1260–1264 (2011)
32. * Long, L.O., Gomer, J.A., Wong, J.T., Pagano, C.C.: Visual spatial abilities in uninhabited ground vehicle task performance during teleoperation and direct line of sight. Presence: Teleoperators and Virtual Environments, vol. 20, no. 5, pp. 466–479 (2011)
33. * Blitch, J.G., Bauder, C.J., Gutzwiller, R.S., Clegg, B.: Correlations of spatial orientation with simulation based robot operator training. In: 4th International Conference on Applied Human Factors and Ergonomics (AHFE), San Francisco CA, vol. 424 (2012)
34. * Chen, J.Y., Barnes, M.J.: Supervisory control of multiple robots: effects of imperfect automation and individual differences. Hum. Factors **54**(2), 157–174 (2012)
35. * Chen, J.Y.C., Barnes, M.J.: Supervisory control of multiple robots in dynamic tasking environments. Ergonomics **55**(9), 1043–1058 (2012)
36. * Fincannon, T., Jentsch, F., Sellers, B., Keebler, J.R.: Beyond spatial ability: examining the impact of multiple individual differences in a perception by proxy framework. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 127–128 (2012)

37. * Barber, D.J., Reinerman-Jones, L.E., Matthews, G.: Toward a tactile language for human–robot interaction: two studies of tacton learning and performance. Hum. Factors **57**(3), 471–490 (2015)
38. * Abich, J., Barber, D.J.: The impact of human robot multimodal communication on mental workload, usability preference, and expectations of robot behavior. J. Multimodal User Interfaces **11**(2), 211–225 (2017)
39. * Reinerman-Jones, L., Barber, D., Szalma, J., Hancock, P.: Human interaction with robotic systems: performance and workload evaluations. Ergonomics **60**(10), 1351–1368 (2017)
40. * Wright, J.L., Chen, J.Y., Barnes, M.J.: Human–automation interaction for multiple robot control: the effect of varying automation assistance and individual differences on operator performance. Ergonomics 1–13 (2018)
41. Fincannon, T., Evans, A., Jentsch, F., Keebler, J.: Dimensions of spatial ability and their influence on performance with unmanned systems. Hum. Factors Defense: Hum. Factors Combat Identification, 67–81 (2010)

* Included in literature review

Classification of Safety-Relevant Activities by Using Visual Scan Pattern in Airport Control Operations

Lothar Meyer[1](✉), Åsa Svensson[2], Maximilian Peukert[1], Sven Malmberg Luengo[2], Jonas Lundberg[2], and Billy Josefsson[1]

[1] LFV Air Navigation Services of Sweden, Research and Innovation,
Hospitalsgatan 30, Norrköping, Sweden
lothar.meyer@lfv.se

[2] Department of Science and Technology, Linköping University,
Norrköping, Sweden

Abstract. Digital remote tower technologies rely on a video presentation for providing safety-relevant information to the tower controller. The quality of the video presentation might affect visual capabilities of the tower controller to perceive all information needed for decision making when changing from conventional to remote tower control. For investigating possible implications on safety-critical activities, we created a first baseline from a conventional tower using a verbal coding method and eye tracking data for classifying periods of visual activities by the related control task. This allows us to identify characteristics in the visual scan patterns using the out-the-window view. The presented proof-of-concept study comprises 12 approach situation samples and three tower controllers in a field study. We found group-specific and individual-specific visual scan patterns that are characteristic activities for undertaking certain control tasks. During visual search for establishing visual contact, all controllers exhibit visual scan patterns forming a triangle consisting of runway, airspace and radar. Individual characteristics were found in the timing and frequency of fixating the areas of the triangle. Also, the times fixating instruments and the out-the-window view are found to be individual characteristic. The findings provide insights into characteristics of the tower controllers that are appropriate for a later comparison with a subsequent analysis of the digital remote tower.

Keywords: Remote tower control · Safety · Visual scan patterns · Eye tracking · Decision-making

1 Introduction

LFV, the Swedish Air Navigation Service Provider, focuses on the digitalization of its services and in particular the deployment of digital air traffic control services for small and medium size airports. The Swedish airports Örnsköldsvik and Sundsvall were put into operations remotely from the remote tower center (RTC) in Sundsvall in the year

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 381–395, 2019.
https://doi.org/10.1007/978-3-030-22507-0_30

2015. The airports Linköping, Malmö, Umeå, Östersund and Kiruna are following in the scope of an ongoing implementation program.

It is then important to investigate the possible influences of a digitalized tower working position on tower controller's behavior of sensing and recognizing safety-relevant visual information for decision making after the transition from conventional tower[1]. The purpose of this paper is to conduct a proof-of-concept evaluation that shall test and assess a new method of classifying visual scan activities through a case study of the visual scan patterns of three air traffic controllers at one airport. Our approach is the analysis of a baseline of the conventional tower that identifies visual scan patterns that are characteristic for the tower controller's work. The identified characteristics shall be used for a subsequent comparison with the digital remote tower environment. We use episode analysis, involving an area-of-interest (AoI) analysis, and so called dwell-time-share diagrams that are specifically developed by us to identify characteristics in the form of reoccurring visual scan patterns and include the out-the-window (OTW) view as the primary information source.

The remote provision of control, information, weather observation and alerting services relies primarily on the video presentation that substitutes the conventional OTW view. The video presentation is enhanced by automatic assistance systems in an integrated platform solution from the industry (SAAB) such as wind sensors, cloud ceiling and weather radar that is used to support the weather observation. Nevertheless, the tower controller's capability to monitor and assess the conditions in the control zone and on the runway is safety-relevant for decision-making. Necessary actions for separating aircraft, such as instrument flight rule (IFR) and visual flight rule (VFR) based movements or recognizing runway conditions during landing and departing situations rely on the tower controller's capability to visually search, find and assess cues using the visual information provided.

A considerable number of research studies addressed related behavior phenomenon by investigating the visual activities in tower control, of which some are briefly presented here. A list of 28 "visual features" was identified by Ellis and Liston [1] from discussions with 24 controllers. At the example of aircraft landing deceleration on the runway, visual velocities and features of anticipating the aircraft were analyzed on the ability of the tower controller to perceive the speed by the visual change. According to the results, tower controller's ability to judge the change of speed is a learned viewing strategy. Complementary to this study, the visual cues perceived by the tower controller are investigated by means of a questionnaire and seven tower controllers [2]. A list of OTW-relevant visual cues was ranked according to the range of perceptibility and importance with the cue "vehicle on maneuvering area" as most important to detect.

An empiric study on the use of the out the window view revealed that the tower controllers identify aircrafts visually for verifying the information provided by the flight strips [3]. Additionally, the airport is monitored occasionally in order to potentially permit an immediate reaction to unexpected events. The sequence of scanning working instruments and areas of interest of the OTW was investigated by using an

[1] This study was funded by the Swedish Transport Administration and the LFV Air Navigation Services of Sweden through the project DIGIT.

episode analysis [4]. The comparison of the tower controller's work patterns regarding system interaction in a multi and single airport working environment revealed an individual variance between tower controllers and the interdependency between work patterns, the design of the environment and the use of implicit communication. The findings reveal the existence of visual working patterns that are characteristic to the individual tower controller and that depend on characteristics of the operational context such as weather and traffic.

Possibly, safety-relevant effects on the working behavior of tower controllers may arise from the fact that the video and visualization technique affects the "in situ" perceived picture compared to a conventional tower. These implications might arise from the fact that the video presentation bases on camera and visualization technology that is still state-of-the-art but nevertheless a reproduction. The design of the video presentation equipment tends to be dominated by questions about display resolution minima where 85% of the population is able to discriminate visually 1 arcsec^{-1} [2]. Comparing camera and human visual capabilities fairly, the range of aspects is more diverse from which two are presented briefly. Exemplarily, the human eye is able to see a huge range of intensities, from daylight levels of around 10^8 cd/m^2 to night luminances of approximately 10^{-6} cd/m^2 [5]. It is capable of working in visual environments with a large luminance range due to a process called "adaptation". At the retina level, eye adaptation is highly localized allowing us to see both dark and bright regions in the same high dynamic range environment. The capability to detect motion by the human eye is performed by amacrine cell that reports salient features of the visual world to the brain [6]. This is an important feature of the peripheral vision that allows the tower controller to keep track of movements on and around the runway including the instantaneous detection of non-authorized movements. Taking into account the findings on controller's work pattern, these are indispensable capabilities of collecting visual evidence for building up situational awareness and decision making in a safety-critical work environment.

With a view on the forthcoming transition process to digitalized remote towers, the direct visual contact of the human eye to the operational environment is substituted by a video presentation. This hence changes the physical origin of visual stimulation. An operational relevance might result from the circumstance that the substitution affects the mentioned capabilities of the tower controller. This possibly affects activities of searching and establishing visual contact to operational-relevant objects in time. A safety-relevant question arises as the early identification of threatening situations such as the runway incursion relies on the timely provision of all visual cues under consideration of the physiologic-visual capabilities of the human.

Our first step of investigating possible implications of the video presentation is to create a baseline that consists of characteristics of scanning behavior for a later comparison with the digital remote tower which we plan for this year 2019. Accounting for the diversity of characteristic scanning behavior [4], we distinguish two key areas that may be subject of implications when changing to a video presentation:

- Group-specific characteristics of visual scan patterns that tower controllers share
- Individual characteristic of a certain tower controller (visual signature).

Both points are interrelated since visual signatures and group-specific scanning behavior exclude each other implicitly.

For the baseline in the conventional tower, we identify characteristics consisting of a systematic and repeated sequence and related timing of the scanning pattern during a specific activity of the controller. The identification shall then succeed by comparing gaze data samples of approach situation samples from tower controllers in live operations. The approach of an aircraft, including final approach and landing, is a high-risk situation in which 49% of the accidents in commercial aviation occur [7]. All operational processes on and around the runway rely on the complete understanding and situational awareness of the controller. Possible erroneous judgments might be caused in incomplete or corrupted scanning patterns by the controller.

Equal conditions of comparison are an essential prerequisite for identification that requires distinguishing and classifying the traffic situation, weather and the activity. The latter refers to the intended task as defined for the tower controller that is assigned to follow the rules as defined by ICAO doc. 4444 PANS-ATM [8]. The intention to undertake a certain control task is a key feature of explaining the variance of the actual scanning activity (based on empirical observations). This is important due to the trained ability of the controller to handle multiple tasks at a time that are serialized by switching the task according to the current demand [9]. By such a task-related classification of the activity, we expect to distinguish and identify even small features of characteristics in the scanning patterns since they feature the same intention of the tower controller.

In the scope of this approach, we present here the results of a proof-of-concept study that has the following objectives:

- Evaluating the method of classifying activities of equal intention
- Identification of the baseline characteristics in the conventional tower.

In the following, we present the setup of the observation study for collecting eye tracking data in live operations. Further, we introduce our verbal coding method that is used to distinguish intention and the related activities for understanding the visual scan patterns. For analysis we use the dwell-time share diagrams for comparing the scan patterns observed. The discussion highlights aspects of the results such as the conditions of recordings and the found characteristics of the scan patterns. Finally, we conclude the major statements possible on the basis of the results gained so far.

2 Method

2.1 Observation Study

The field study was conducted at the “SAAB” Linköping Tower during two days and involving three tower controllers. Figure 1 provides an overview of the tower working position, including areas of interest (AoI) (see Sect. 2.3 Episode Analysis for further details of the AoIs). Controller A is 30 years old, with an operational experience of 2.5 years, B is 29 years, with an operational experience of 4 years and C is 42 years with 22 years operational experience. The tower environment was selected due to the fact that Linköping Tower is planned to be remotely controlled in spring 2019. Eye-point-of-gaze (an indicator of visual attention) in the conventional tower was measured by means of a Tobii Pro Glasses eye-tracking equipment, with three tower controllers. The Tobii glasses provided data of eye gaze movements with a sampling rate of 50 Hz extended by audio and video captures of the scene. The use of eye tracking-related terminology refers to the definitions made in [7]. The analysis was performed using Tobii Pro Lab 1.76 and specifically programmed tools on the basis of Java.

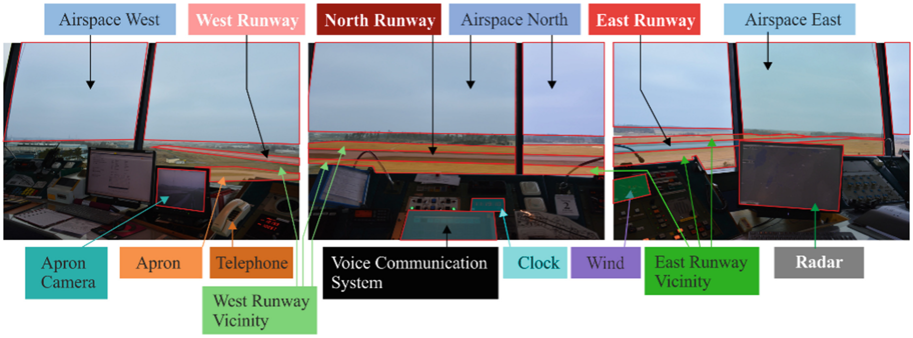


Fig. 1. Areas of interest of Linköping Tower

2.2 Verbal Coding

A usual practice in eye gaze analysis is to use radio voice communications for disclosing intent and thus to classify the observed activities (e.g. empirical observations of the controller’s work). In contrast, the use of the window view is in the majority of the situations featured by radio silence and thus does not provide sufficient cues for concluding on the actual intention and classification of the related activities.

A key requirement in our approach was to relate the intent of the tower controllers to the observed visual activities. The aim was to classify episodes of activities and thus to identify similar situations of using the window view. Our solution to the issue of identifying intent was to conduct an “in situ” verbal coding that extends the recording by an active support of the tower controller. Beside the task to provide tower control services, the controller was advised to utter clearly a code to indicate the current visual activity. A list of verbal codes was evaluated by the tower controllers and reduced to a basic and simple set that all refer to the use of the out-the-window view:

- “Check”: The controller checks the runway for obstacle clearance.
- “Birds”: The controller checks for birds on or around the runway.

- “Search”: The controller search for expected approaching aircraft in the airspace.
- “Contact”: The controller establishes visual contact to the expected approaching aircraft.

The verbal coding provides an important subjective reference time of the true event of establishing visual contact. The chosen approach thus permits for narrowing the selected time period of recordings down to the desired search activities.

2.3 Episode Analysis

As mentioned in the introduction, our analysis approaches the identification of characteristics in the sequence and timing of scanning visually the working environment. Therefore, the chosen analysis method applied is the “episode analysis” which allows bigger audio and video data sets to be divided into shorter episodes, or sub-episodes, for in-depth transcriptions [4, 10, 11].

To narrow down the times of interest, we divide the audio, eye-gaze video into episodes of interest for in-depth description. The episodes of interests are arrivals of aircraft to the airport, from the initial call of entering the control zone till the touch-down. The purpose of this analysis is to identify the times of activities that are related to the visual search for the expected aircraft and other OTW-related activities. The visual search might be embedded within the task of handling an approaching aircraft beside activities such as note taking on the flight strip or communications with the aircraft. Thus, the episodes help us to determine the periods of visual search and to predict the intention of the tower controller independent of the verbal coding.

To determine visual scan patterns, areas of interest were defined as shown in Fig. 1. It shows the view over the runway from the working position and surrounding equipment such as the radar, clock and an additional video view. The 15 AoIs are complemented by the flight strip-AoI and the vicinity of the runways divided into a lower and an upper part each (18 AoIs in total). On the basis of the area-of-interests and episodes, the related dwell times are calculated indicating the share of attention over the area-of-interests. The dwell-time-share diagrams developed for this purpose highlight the visual scan pattern of the three tower controllers during approach situations in an easy understandable way.

3 Results

The analysis bases on eight hours of eye tracking recordings from the three tower controllers. The recordings were conducted in a period of August and September 2017 as well as February 2018 in a time between 8 am and 2 pm when higher volumes of traffic were expected including VFR and IFR. The weather conditions had a visibility above 12 km with scattered till broken clouds during all recordings. The runway in use was 29, meaning aircraft approached from the east side on the controller’s right. From all recorded situations, four samples per tower controller were chosen for the episode analysis containing one approach situation each.

3.1 Dwell Time Share

The dwell time diagrams (Figs. 2, 3 and 4) show the temporal distribution of fixations on the 18 specifically defined AoIs indicated by the color of surface. We defined a fixation with a minimum duration of 60 ms not exceeding a speed of 30 deg./sec.

The diagrams cover the chosen episode from 4 min before the touch down (or touch and go) till 30 s after the touch down (or the touch and go). 4 min was chosen due to the time of the aircraft having entered the control zone. This provides a complete picture of the working context including the entire approach situation. The graph distinguished fixations on the OTW as surfaces on the upside of the coordinate axis whereas head down fixations lie on the downside. The AoI-states are originally binary distributed resulting in the majority of the cases in a highly fragmented picture of the context. Therefore, we applied a smooth filter by using a sliding window with a size of 2000 ms (symmetric range 1000 ms). This allows us to filter out the long term work pattern by reducing the noise of fragmented AoI-fixation that is caused by high frequency changes. Complementary, we use a fixation time metric that indicates the relative mean length of a fixation within the sliding window by a black graph. The graph indicates situations in which AoIs were scanned more intensively than others. An example with low fixation times is the runway check where the controllers slips visually across the runway using saccadic eye movements. This is in contrast to the use of the radar screen that exhibits often times long fixations times.

The diagrams have an additional label on the upper axes indicating the times where the tower controller uses the verbal codes. These are complemented by event labels such as the moment of giving a clearance to the aircraft, landing as well as touch and go events. Some graphs are left truncated due to operational limitations of initiating and calibrating the eye tracking device while providing control services.

The diagram shows the labels “search”, “contact” as well as “check” in all the approach samples. The code “birds” was used 8 times by 2 of 3 persons. All “search” and “contact” codes were used during periods of fixating the OTW. Corresponding, the code “check” was used while fixating the runway. The code “bird” gave mixed results as it was not clearly associated with a certain area rather it can only be stated that the point of fixation was on the runway or the vicinity of the runway.

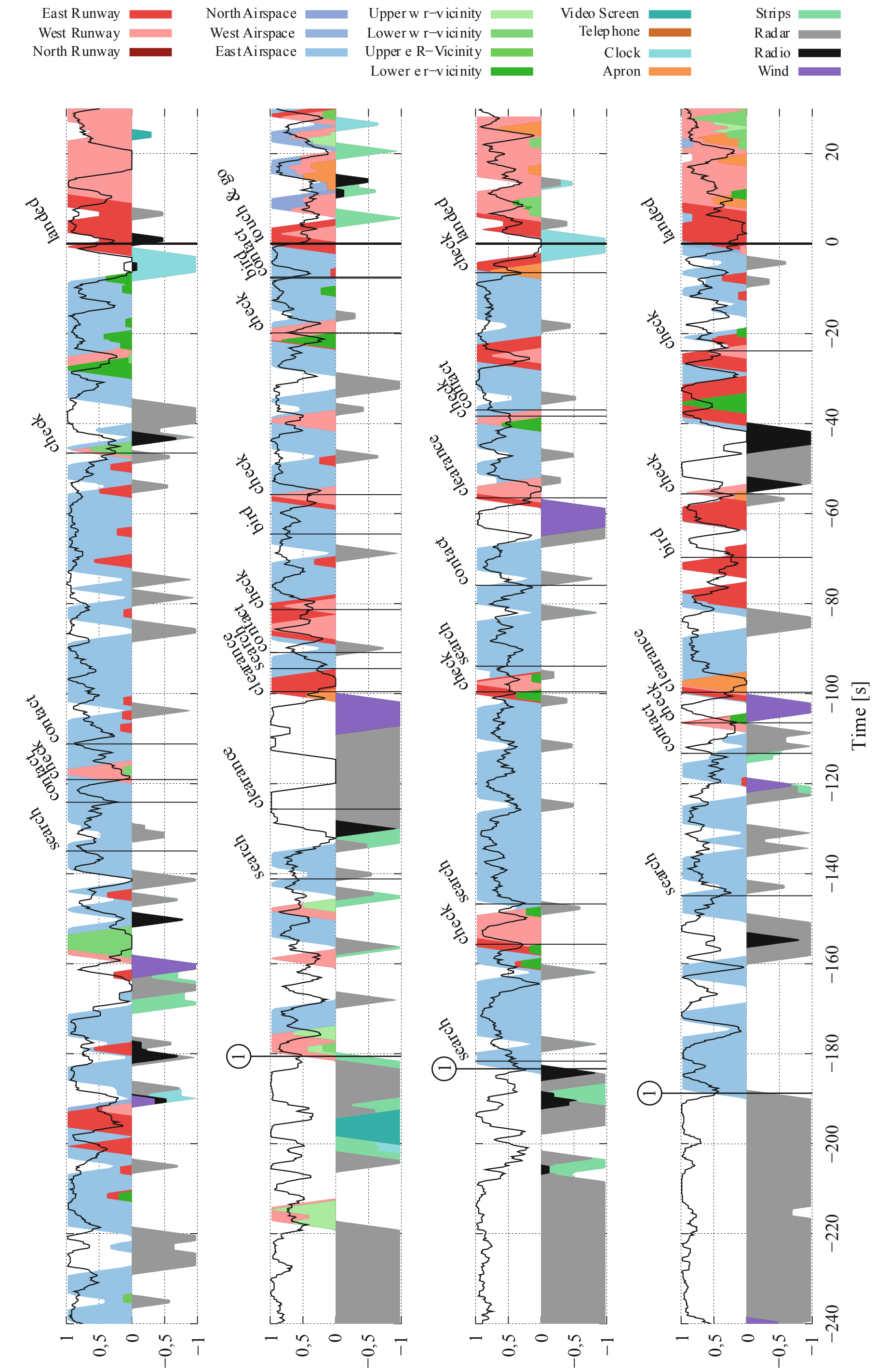


Fig. 2. Controller A

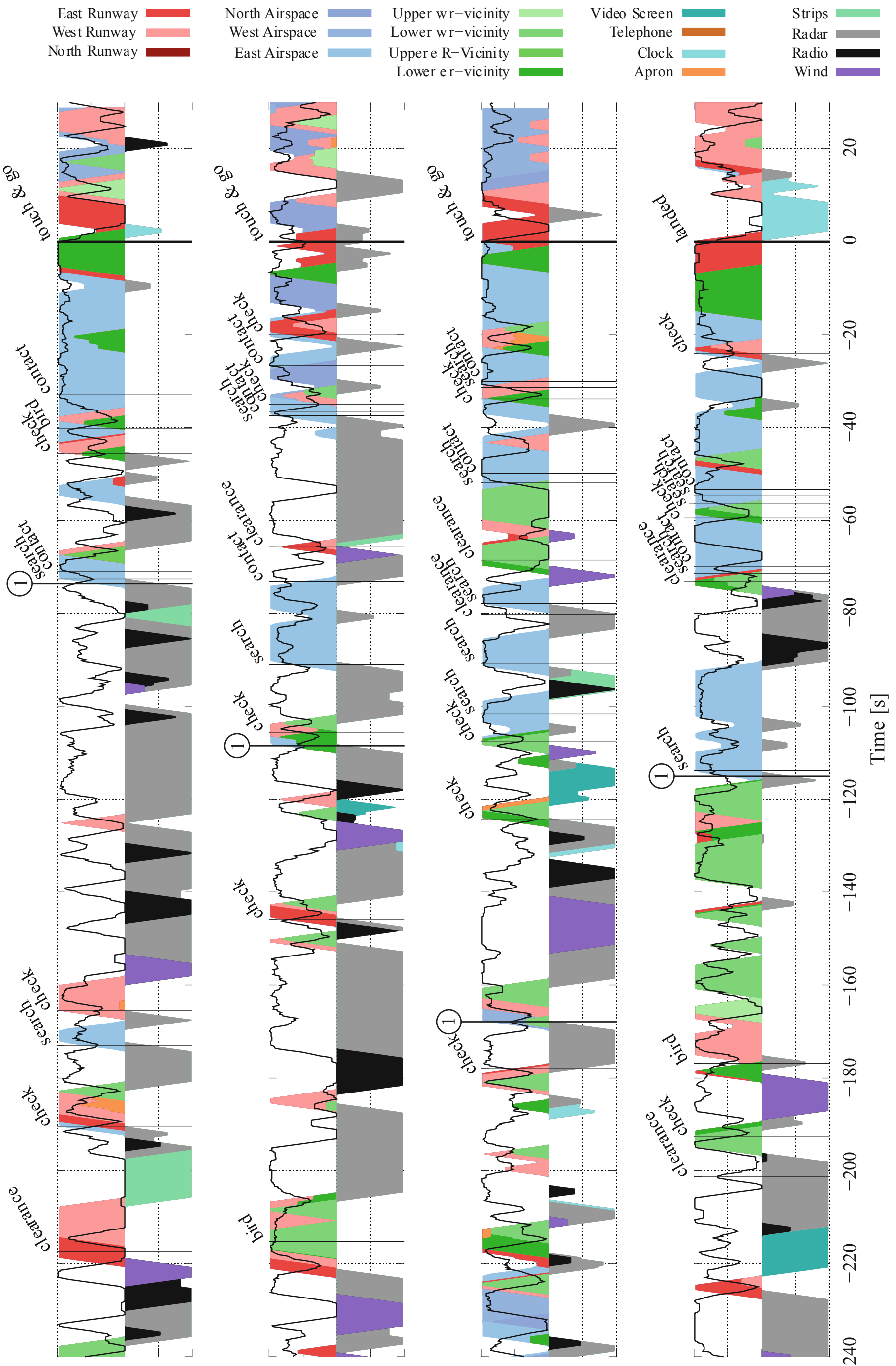


Fig. 3. Controller B

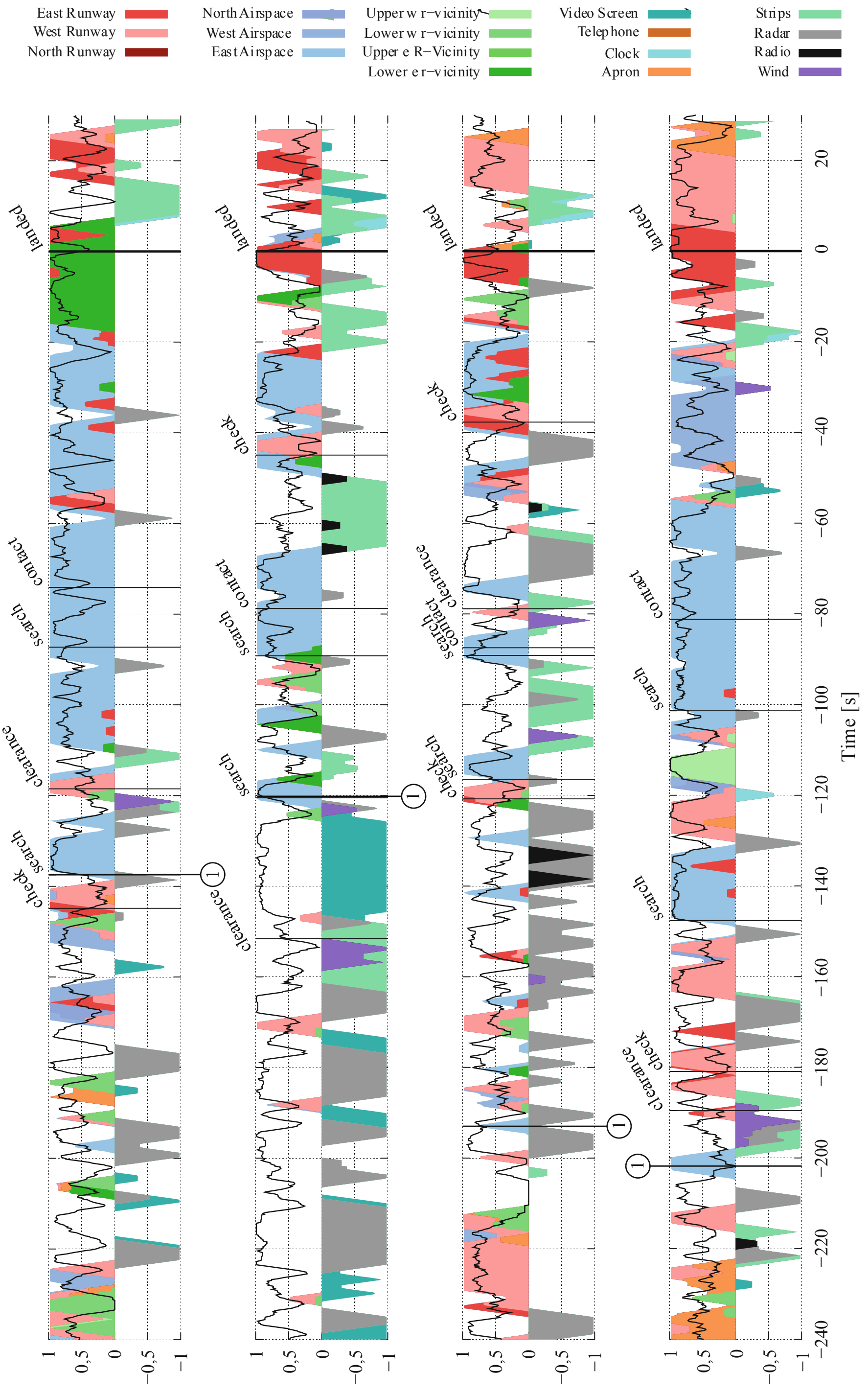


Fig. 4. Controller C

3.2 Statistics for the Areas of Interests and Verbal Coding

Times of Verbal Codes

Looking at the analysis of the times of verbal coding, the establishment of first visual contact had a mean of 81.9 s (SD 19.21 s) before the moment of touchdown. Controller A distinguished from the other controllers with an early establishment of visual contact in sample 1 (124.1 s) and sample 4 (113.2 s). Controller A showed also the highest dwell times at the east airspace with between 18.6 and 52.1% within the episode (Table 1). Controller B showed the latest establishment of first visual contact with times between 49.8 s and 73.1 s. The mean time between calling out "search" and "contact" was 10.2 s (SD 9.6 s).

Table 1. Dwell times of selected AoIs in percent

	A				*B*				*C*			
Sample	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*
Radar	14.5	25.8	30.4	41.4	18.3	42.7	14.4	18.7	13.0	23.9	21.0	8.9
Radio	2.1	0.7	1.0	2.6	4.8	5.6	3.8	2.3	0.4	0.6	1.3	0.2
Wind sensor	1.4	1.6	2.0	1.9	1.3	5.7	3.9	3.3	0.6	1.7	1.1	1.4
East RWY	8.6	3.7	4.5	10.3	2.6	4.5	2.7	7.0	6.0	4.5	5.8	6.2
West RWY	11.4	10.5	11.5	7.0	6.0	7.8	5.3	10.9	8.2	6.9	16.2	24.0
North RWY	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
East airspace	52.1	18.6	41.6	23.4	6.1	8.4	14.6	21.8	32.4	15.7	11.5	21.7
West airspace	0.3	1.0	0.0	0.6	1.1	0.2	5.6	0.0	3.7	0.6	2.0	5.5
North airspace	0.0	0.7	0.0	0.0	0.0	8.1	3.2	0.0	0.7	0.0	0.2	1.2
Flightstrip	1.3	4.6	1.5	0.36	2.0	0.6	0.6	0.0	7.4	13.6	7.2	4.7

Area of Interest statistics

Comparing the overall mean of the dwell times on the basis of Table 1, the use of the radar exhibits the highest overall share of fixations with 22.8% followed by the east airspace (22.3%) and the west runway (10.5%). This so called AoI-"triangle" consists of the three most used AoIs that are balanced individually by the controllers in terms of the total amount of attention as well as quality the timing of switching in between these AoIs. According to Table 2, Controller A shows in the episodes a focus on the OTW while Controller B has a rather radar dominated pattern. Controller C has a tradeoff with an increased tendency to the runway compared to controller A.

Table 2. Tradeoff runway, airspace east and radar per controller

	A	*B*	*C*
RWY	16.9	11.7	19.4
Airspace east	33.9	12.7	20.3
Radar	28.1	23.5	16.7

4 Discussion

The results show the visual work pattern of 12 approach samples and three tower controllers by means of the dwell-time-share diagrams. The eye gaze data contains the observation of how the control tasks are in facto executed in a conventional tower. The 12 samples cover, therefore, an enormous amount of information as gaze data that is in raw form. The raw form of such data is neither readable nor understandable for investigating the work pattern of tower controllers using the OTW. In this regard, both the AoI analysis of the episodes and the application of a sliding window filter helped to structure the gaze data and thus to increase readability and understandability. The analysis was additionally labelled with contextual information of the operational situation and the intention of the tower controllers while using the OTW. The resulting dwell-time-share-diagrams provides on overview of the work patterns that provides the best prerequisites for identifying characteristics in the sequence of AoIs and the related dwell time.

Based on the work patterns of the dwell time-share diagrams, the verbal coding and the statistics, we found several indications of group-specific systematic working shared by the three tower controllers.

- As explained initially, the intention to undertake a certain control task is a key feature of explaining the variance of the actual scanning activity (based on empirical observations). Within the episodes, we were able to distinguish the periods of control tasks and related intentions that we define as following:
  - *Entry of aircraft into control zone:* While the aircraft is approaching the airport, after entered the control zone, but still far enough from the tower to be visually noticeable in the sky, the intentions by the controller are to plan and prioritize the runway usage using the flight strip system. However, there were also several short periods of using the OTW that we explain by a demand for scanning the environment for indications that helps to anticipate upcoming runway usage in terms of expected departure and arrival movement. On the ground side, the monitoring activities include aircraft on the apron preparing the departure. These activities might aim on visual cues such as the boarding or refueling using the apron camera and the apron sight. On the air side, VFR traffic is observed that is located in the controlled airspace using the radar or the OTW. In general, the predominant visual sources are AoIs located on the instrument panel.
  - *Visual search for approaching aircraft:* The initiation of visual search activities is indicatable by increased dwell times on the approach airspace. Most likely, the moment of initiation is triggered by the traffic situation presented by the radar. The moment when the controller switch attention from the radar to the approach airspace is indicated in dwell-time-share diagrams (Figs. 2, 3 and 4) by the flag labeled as “1”. The moment is in 10 of 12 cases accompanied by a direct change from radar to approach airspace. The triangle of radar, approach airspace and runway checks is established in the following. The runway checks are in most cases applied after the search at the approach airspace.
  - *Established first visual contact:* By statistical analysis of the verbal codes times, the time of visual establishment was determined at a mean of 81.9 s before the

  touch down event shows a rather low 19.2 s standard deviation. The landing aircraft types varied from a small P28A till an Embraer 190 with quiet different dimensions of the visual cross-section and thus different prerequisites of detecting the body. An explanation for the nevertheless homogenous time of establishing first visual contact might be the correlation between size of aircraft and its speed during approach. Smaller aircraft are detected closer to the runway which is counterbalanced by the lower speed. After successful establishment, the pattern of the triangle remains. The focus might shift in some cases to an intensified monitoring of the runway vicinity, including the taxiways to the runway indicated by "apron" and "lower west runway vicinity".
  - *Full stop landing:* The landing event is indicated by the fixation of the clock and the flight strip, the controller notes the time of landing. The landed aircraft is visually followed on the runway while monitoring the taxiways to the runway.
  - *Touch & go:* The touch and go is accompanied by following the aircraft visually at the airspace west. The most likely explanation for this is to see the aircrafts turning into a right aerodrome circuit as usual cleared at this airport.
- The runway check is indicated by short fixation duration and a rather high number of saccades during the scan.
- The landing clearance was announced in 11 out of 12 cases by the fixation of the wind info.

Within the work pattern, the dwell time-share diagrams (Figs. 2, 3 and 4) indicate several observations that point on the individual signatures on how the task is applied. Exemplarily, controller A and C focused mainly on the airspace for an early establishment of visual contact and embedded short episodes of checking the runway for obstacles. In contrast, the controller B tends to use the radar instead of following the aircraft visually before establishing visual contact and planned already for the next traffic movements at the same time. Distinctions in the efforts on establishing visual contact to aircraft on final are clearly indicatable by the time spend on the approach sector. This corresponds to an individual trade-off between planning and prioritization of monitoring movements. More specifically, the controllers used the flight strips, radar and clock, for planning activities on the first hand and the separation activities on the other hand, involving the runway, position fetching on radar and the window view, as well as occasional monitoring of unexpected obstacles on or nearby the runway. The following features might summarize these distinctive features of the three tower controllers that is considered as characteristic and systematic for the individual:

- The timing of the task switching
- The tradeoff of directing the visual attention between the runway, approach airspace and radar
- The runway checks involving the visual check of the taxiways and runway holding points individually
- The bird check.

The discussion relies on 12 samples of a field study that was conducted under rather constant operational conditions in terms of weather and air traffic movements. Nevertheless, the variability of the conditions does not allow for a generalization of the results since the sample size does not account for the related complexity of the operations. This concerns especially the confounding effect of other air traffic on ground or in the control zone as well as planned activities of the airport operator that might influence the activities. The results are disturbed by the chance that the controllers did actually not verbalized the current intention at all opportunities available during the recordings. Rather, the visual scan pattern that can be related to a certain intention provides a template that allows for identifying similar periods in the episode.

5 Conclusion

The paper presents a proof-of-concept study that shall test and assess our method of classifying visual scan activities. The verbal coding helped us to understand and relate the observed visual scan pattern to the intention that allows us to identify the times of switching between the tasks during the approach. By this, we were able to classify the periods of executing certain control tasks and to compare them for identifying characteristics of the tower controllers. The dwell-time-share diagrams showed clear distinctions between tower controller's scan pattern of gathering activities within the chosen periods. The differences were shown in terms of time and efforts spent on specific control activities and the related sequences of focusing on specific visual cues. This concerns in particular the tactics of the controllers to search (visually) for the aircraft in the controlled airspace and on final.

The results show the success of our visual scan pattern analysis method to classify activities while executing a control tasks and to identify differences between controllers. The method, which will be used to proof safety-relevant implications during the transition to digital tower control operations, is however still under development. The focus of future research activities is set, therefore, on the evaluation of robust metrics that shall indicate the statistical significance of the signatures found so far.

References

1. Ellis, S.R., Liston, D.B.: Visual features used by airport tower controllers: some implications for the design of remote or virtual towers. In: Fürstenau, N. (ed.) Virtual and Remote Control Tower. RTA, pp. 21–51. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28719-5_2
2. van Schaik, F.J., Roessingh, J.J.M., Lindqvist, G., Fält, K.: Detection and recognition for remote tower operations. In: Fürstenau, N. (ed.) Virtual and Remote Control Tower. RTA, pp. 53–65. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28719-5_3
3. Pinska, E.: Warsaw Tower Observations. Eurocontrol Experimental Center, Paris (2007)
4. Svensson, Å., Forsell, C., Johansson, J., Lundberg, J.: Analysis of work patterns as a foundation for human-automation communication in multiple remote towers. In: Proceedings of the Twelfth USA/Europe Air Traffic Management Research and Development Seminar, Seattle, USA (2017)
5. Ledda, P., Santos, L.P., Chalmers, A.: A local model of eye adaptation for high dynamic range images. In: Proceedings of the 3rd International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa, Stellenbosch, South Africa (2004)

6. T. Kim, Soto, F., Kerschensteiner, D.: An excitatory amacrine cell detects object motion and provides feature-selective input to ganglion cells in the mouse retina. eLife **4** (2015)
7. Boeing: Statistical summary of commercial jet airplane accidents worldwide operations 1959 – 2017, October 2018. www.boeing.com/news/techissues/pdf/statsum.pdf. Accessed 15 Feb 2019
8. ICAO: Doc 4444 – PANS-ATM. Procedures for navigation services – air traffic management, 16th edn. ICAO, Montreal, CA (2016)
9. Fischer, R., Plessow, F.: Efficient multitasking: parallel versus serial processing of multiple tasks. Frontiers Psychol. **6**, 1366 (2015)
10. Korolija, N., Linell, P.: Episodes: coding and analyzing coherence in multiparty conversation. Linguistics **34**(4), 799–832 (1996)
11. Rankin, A., Dahlbäck, N., Lundberg, J.: A case study of factor influencing role improvisation in crisis response teams. Cogn. Technol. Work **15**(1), 79–93 (2011)
12. Seetzen, H., et al.: High dynamic range display systems. ACM Trans. Graph. **23**(3), 760–768 (2004)

How Fire Risk Perception Impacts Evacuation Behavior: A Review of the Literature

Hua Qin[1,2] and Xiaotong Gao[1,2](✉)

[1] Department of Industrial Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
qinhua@bucea.edu.cn, 1093881839@qq.com

[2] Beijing Engineering Research Center of Monitoring for Construction Safety, Beijing 100044, China

Abstract. Objective: The objective of this paper is to review the literature on the processes by which individuals respond to fires in order to identify the decision-making process between fire risk perception and evacuation behaviors.

Method: According to evacuation timelines, a conceptual framework is used to identify the mechanisms through which fire cues, the characteristics of the building in which the fire occurs, and the demography, personality, fire experience and training of the individuals may be interpreted as fire risks. The fire risk perception is used as precondition impacting on decision-making and human behaviors. The relevant literature has been searched through electronic databases, journals, and consultation with key informants.

Results: People respond differently to various perceived fire cues. Actions depend on the cues perceived, the interpretation of the situation, and the subsequent decisions taken. Occupants act based on these decisions, but new information can cause them to discard previous actions and begin new processes. In addition, many of the actual behaviors of occupants in fatal fires differ from occupants' response performance models.

Conclusions: The fire cues perceived by people would be interpreted as safe or risk.

This interpretation process is affected by several factors, including the features of the fire cues, the architecture of the building in which the fire occurs and personal characteristics. The interpretation also impacts importantly on the decision-making and responding behaviors.

Keywords: Risk perception · Building fire · Evacuation behaviors

1 Introduction

In many fires, research on fire injuries and deaths shows that over two-thirds of the injured and over half of the dead in building fires could have evacuated. But these people delay their safety inside the building [1, 2]. A solution to this problem is to make sense a comprehensive and validated theory on human behavior during evacuation from building fires. If the persons perceive a fire cue and think that is a risk, they will intend to evacuate. This decision indicates that fire risk perception impact potentially on persons' responds.

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 396–409, 2019.
https://doi.org/10.1007/978-3-030-22507-0_31

This paper reviews the processes by which individuals respond to fires. These responses begin with the perception of fire cues. Fire cues may be interpreted as either safe or indicative of a fire risk, depending on the features of the cues, the characteristics of the building in which the fire occurs, and the demography, personality, fire experience and training of the individuals. Each individual incorporates these factors in a decision-making process to identify necessary protective actions and to form an adaptive plan or strategy. After decision-making, the individuals carry out the actions.

2 Method

2.1 Literature Search

For the purpose of the present literature review, we followed the steps for a systematic literature review. Firstly, we searched related database such as "Web of Science", "Google Scholar". The databases searched were about "building fire", "risk perception", "decision making", "evacuation behavior", "pre-movement time" and so on. Then the in-depth review took place. After that, main questions were proposed: the process of risk perception, the factors of impacting risk perception, the process of risk perception impacting on decision-making and so on. Finally, the literature was included if it was relevant to the topic.

2.2 Frame of the Review

The behavior of occupants during building fires has been shown to affect survival rates significantly. Survival probabilities are largely determined by occupants' responses during the fire [3]. Therefore, the relationships between the evacuation processes, fire development (including ignition, the initial period, the fire development period, the flourishing period and dying out) and safety should be investigated. It has been shown that after occupants have determined the fire status of a building, they estimate the risk involved, make decisions, respond to the fire and evacuate the building [4–6] (Fig. 1). If the required safe egress time (RSET) is less than the available safe egress time (ASET), occupants can evacuate safely. The pre-movement time critically affects the RSET. However, research has shown that the pre-movement time in actual evacuations can last from five minutes to over 25 min [7–9]. Longer pre-movement times represent greater levels of risk to the occupants. However, most evacuation models primarily focus on the purposeful evacuation of occupants and do not consider the perceptual and cognitive processes related to decision-making for real-life evacuation, which may delay evacuation time. Therefore, it is important to characterize how the perception of fire risk influences human behavior.

IG: ignition (start of the fire); AL: alarm (sounding of the alarm); RC: recognition (occupants perceive the alarm); RS: response (occupants respond to the call to evacuate); DD: dangerous (the fire or its products are deadly to the occupants); ET: extinguished (the fire is extinguished); ASET: available safe egress time; RSET: required safe egress time.

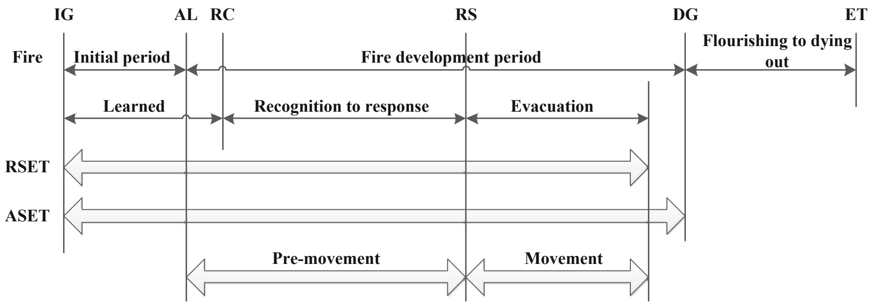


Fig. 1. Evacuation timelines

3 Results

3.1 Fire Risk Perception

Risk Perception in General. To define risk perception, the term "risk" itself must be explained briefly. There are many definitions of risk [10–12], but it is commonly considered to be the likelihood that persons will experience the effect of danger [13]. The various definitions of risk share one common element, that is, the distinction between reality and possibility [14, 15].

Risk perception is the subjective judgment by which people recognize the features of accidents and the severity of risks [14, 16]. The term risk perception emphasizes that risk is assessed based on experience as well as available information. In addition to the features of the risks themselves, personal experience, memory and other socio-demographic attributes or psychological dispositions influence the way people perceive risks, i.e., risk perception is a social construct [17, 18]. People have different comfort levels and adjust the riskiness of their behavior correspondingly [19]. Therefore, risk perception is not a constant but varies between individuals and contexts. Individual risk perception depends on the environment, as well as on the likelihood of the risky outcome and the individual's concerns about this outcome [20].

Perceiving Fire Cues. Afire risk can be defined as the probability that a fire causes casualties and/or property damage [21]. Occupants will evacuate only if they perceive a situation as dangerous. The situations perceived as normal or risky results in different activities. Therefore, ensuring the correct interpretation of fire risk is essential to an evacuation plan [22].

In a fire, the building occupants receive both physical and social external cues. Physical cues include features of the fire itself, such as flames, smoke, and explosions, along with fire alarms, such as tone alarms and automatic warnings. Social cues include communication with other people in the same building or outside the building, the actions of other occupants and/or yelling in the building. These cues can be perceived by multiple sensory modalities, including hearing, smelling, seeing and touching [23].

Research has shown that the occupants' characteristics and the nature of the cues affect the occupants' perception of the fire cues [24].

- Previous experience with disasters or fire training increases the probability that occupants perceive fire cues [25, 26]. Other factors, such as limited experience with the environment, perceptual limitations, and age or stress decrease the probability of cue perception [26–30].
- The perception of fire cues is affected by the ignition location, the physical characteristics of the smoke and the number of cues [23, 30–32]. Kobes, Helsloot, De Vries, and Post [33] found that occupants' perception were impacted by fire characteristics, such as the growth rate, the smoke yield, toxicity, and heat. The fire growth rate is a particularly important factor. Many fatal incidents can be attributed to rapid fire development despite the initial perception of fire cues [34]. Researchers have also found that the credibility of the source of a warning message, along with its delivery method, repetition and consistency, also influences occupants' perception [35–37].

After perceiving such cues, occupants become aware of changes in their environment [1, 38].

3.2 Interpretation of Critical Factors and Decision-Making

Interpretation. Individuals perceive cues that imply that the normal situation has been interrupted and disrupted, creating uncertainty: this information must then be interpreted by the individual [38]. Occupants organize the cues into a meaningful framework or story to make sense of their environmental situation. The construction of these frameworks and stories is called interpretation [26, 38–41].

There are generally three types of factors that influence human behavior in the event of a fire: the fire characteristics, the building characteristics and human characteristics [3, 33, 34, 42–45, 97]. The first two factors are external: the nature of the fire itself and the physical environment within which the occupants respond to the fire. The third factor of human nature is an internal factor. These critical factors are presented in Table 1. Previous studies have shown how these factors directly impact response performance; however, it is more accurate to say that these factors influence cognitive processes. The three factors affect how the occupants interpret the fire situation.

A signal is commonly disregarded as a clear indication of danger. However, occupants believe that they are at risk if they smell smoke or toxic gases or see flames and smoke [45]. If occupants are presented with several fire cues or a consistent set of cues, they will interpret the situation as posing a fire risk [22, 35, 46]. Warnings delivered with a tone of urgency are more likely to be interpreted as representing risk [47]. Hypervigilant persons are more likely to interpret emergency cues as dangerous [48, 49]. Occupants with previous experiences of fires or evacuation drills are more likely to define the situation in terms of a fire risk [1, 38]. However, if the occupants are familiar with the evacuation routes or have previously had frequent false fire alarm experiences, they are less likely to interpret the situation as risky [50, 51]. Researchers have also found that social cues, such as calls from friends and screams or evacuation activities of others, promote the interpretation of a situation as a fire risk [52, 53].

Table 1. Critical factors for perceiving a fire situation as presenting a fire risk

Fire characteristics	Human characteristics	Building characteristics
Perceptual cues • Visual Flame, smoke, deflection or collapse of wall or ceiling • Audible Cracking, broken glass, and objects falling • Smelling Smell of burning • Tangible Heat	*Profile* • Gender, age, and family composition • Education • Observation and judgment abilities, mobility, and physical and mental limitations • Culture	*Occupancy* • Occupant density • Building type: office, factory, hospital, hotel, cinema, college or university, and shopping center
Fire growth rate	*Experience* • Fire experience, fire training, and other emergency training • Familiarity with building	*Architecture* • Number of floors • Layout and building shape • Maintenance
Fire alarm • Alarm signal • Voice communication •Others' actions	*Situation* • Alone or with others • Awareness • Physical position • Working, sleeping, eating, and shopping • Stress and time pressure • Others' reactions	*Refuge area* • Complexity of evacuation route and wayfinding • Location of exits and stairwells
	Personality • Influence of others • Leadership • Negativity toward authority	

Perceived cues cause occupants to develop cognitive images of what they imagine is occurring [1]. Some researchers have found that interpretation methods include the recall of previous behavioral scripts, mental simulation and models [26, 54, 55].

Decision-Making. People respond differently to various perceived cues; however, even when presented with the same cues, people are likely to respond in different ways [56]. Actions depend on the cues perceived, the interpretation of the situation, and the subsequent decisions taken. Occupants act based on these decisions, but new information can cause them to discard previous actions and begin new processes [1].

An individual's actions primarily result from a decision-making process. Most evacuation models significantly simplify behavioral processes, either by assigning a delay time before evacuation rather than considering the situational decisions and interactions of the occupants, or by assigning a behavioral itinerary to the occupants [1]. If people cannot answer the questions that emerge in the course of decision-making, they continue to seek information [36], i.e., people continue their original activities if the cues are not perceived as being sufficiently significant to interrupt these activities. Thus, people may not respond quickly to a fire alarm without additional confirmatory cues. Further investigation and research should be conducted to elucidate the additional information required and the forms and timing with which this information should be provided.

Gigerenaer and Selton [57] outline a two-step decision-making phase in which action options are first generated, followed by selecting one of these options. Action options are generated based on interpreting a situation, where searching for options involves mental simulation similar to that involved in developing interpretations [26, 58]. Although occupants are expected to search for a sufficient number of options during the decision-making phase, some researchers have found that occupants usually find very few options. Time pressure, mental resources and training and knowledge of procedures can lead to a deficit of options [26, 57, 59–61]. Two types of choice strategies may be used to select an option. The first is a rational choice strategy, by which persons optimize decision-making by choosing the best of all available options [62, 63]. The second is a satisficing strategy, by which individuals choose the first workable option [57]. Klein [26] hypothesizes that the rational choice strategy is more likely to be adopted by persons trying to optimize a decision, whereas the satisficing strategy is more likely to be used under time pressure, dynamic conditions and other stressors. People make decisions about risk by focusing on explicit cues and information rather than on all of the available facts. People conceive of risk as the result of a likely outcome or the probability that an outcome will actually occur [20].

Therefore, some researches [64–67] believe that risk perception can be understood as a threshold mechanism for evacuation decision-making. That means the evacuation decision-making is “triggered” if the perceived risk becomes unacceptable. Applied to the situation of fires, cues such as the smell of smoke or other people moving to an emergency exit may activate protective behaviors [68, 69].

During fires, people often make decisions according to their own interpretation schemas. However, researchers find that the people communicating with other persons or technology may change their decisions in the interpretation of cues. But potential influence of communication is under exploited at present [70, 71].

3.3 Human Behaviors

Behaviors in the Pre-movement Period. After hearing a fire alarm, occupants always spend a period of time in a non-evacuation activity. This stage is called the pre-movement period [1, 45, 72]. Kobes [3] combined the phases of clue validation and decision-making into this pre-movement period. Incident analyses indicate that pre-movement time and pre-movement behavior play key roles in the evacuation process [73]. This pre-movement time is a delay during which occupants attempt to gather information, alert others in the building, gather their personal belongings, assist or rescue other persons, wait, or fight the fire [1, 74, 75]. Research has shown that pre-movement delays are increased by certain actions such as searching for information or confirming information about an incident [9, 72, 76].

Kuligowski and Hoskins [72] investigate a fire incident and found that the main factors affecting the pre-movement time are the activities undertaken during this period and the initial floor location of the occupants. Another key factor is the information provided to the occupants during the fire. Occupants' past experiences with emergencies and their evacuation knowledge can affect their actions in this period [26, 36]. Researchers have found significant differences between the protective behaviors of women and men during this period. For example, women were more likely to gather their personal belongings, while men were more likely to fight fires or to rescue others such as family members and friends [77–79].

These behaviors show that mental processes and actions involve continuous information-processing and decision-making. During an evacuation, people are often confronted with smoke, toxicity and communication with other occupants. Under these conditions, people may change direction because of breathing difficulties, limited visibility or other reasons [58]. People constantly evaluate their options during an evacuation based on their perception of the following factors: time pressure, perceived safety, implementation barriers and the costs of taking an action [36]. People usually act in similar ways in dense areas such as shopping malls or indoor stadia, because group behavior plays a significant role throughout the evacuation process [80, 81]. Uncertainty before evacuation or other types of survival behavior causes people engage in additional information-seeking until they locate sources or channels that can enable them to make decisions [9, 82]. People do not necessarily progress through the aforementioned stages in order [36].

Evacuation Behaviors. There are three response performances exhibited in surviving a fire [51]: extinguishing the fire, waiting for rescue and evacuation. These strategies can be separated into those conducted during and after the pre-movement period. Research has shown that when visibility is limited, occupants prefer to walk alongside walls or jump out of a building rather than wait to be rescued.

Occupants' walking speed under smoke exposure is also slower than that under normal conditions [58, 83, 84]. Occupants who are familiar with a building prefer to take the stairs over the elevators during evacuation [49]. Occupants who are not familiar with the building exits follow other occupants to the stairs [85].

Individual behaviors in a crowd are always influenced by other persons. Studies of incidents show that most people act as followers. People do not react to a fire alarm until others take action [86]. A set of individuals in the same physical environment is considered to be a crowd [87]: crowd behavior corresponds to the dynamics of entire groups of people resulting from their interactions with the environment. The crowd mind propagates through the crowd by anonymity, contagion and suggestibility. Thus, the intentionality of the individual vanishes in preference to that of the collective [88]. Under some circumstances, individuals in a crowd may not able to conduct a task because their judgment abilities have been degraded to some extent [89]. Although there have been many studies on the effects of crowds, more in-depth research is needed to determine the process and mechanisms by which a crowd impacts individual behaviors and to evaluate the extent of this impact. Does the influence of the crowd increase with the emergency of the situation?

Software development has facilitated the construction of many evacuation models. However, few models have been based on human behaviors, such as escape route selection and information interpretation [33]. The deficiencies in these models can be attributed to the scarcity of relevant and quantitative research data. Various human behaviors have also not been sufficiently understood. The following aspects of human behaviors require further study [34]:

- the response performance and activity patterns of occupants upon hearing fire alarms,
- the occupants' response time and evacuation strategy, and
- wayfinding during evacuation: while most existing studies have focused on architectural construction and building layout, few studies have considered how the layout of the architecture or the building affects decision-making.

Disparity Between Response Performance Models and Actual Behaviors. Both the technical and social aspects of building fire safety policy are based on the paradigm of occupants' response performance; however, many of the actual behaviors of occupants in fatal fires differ from occupants' response performance models. The disparity between response performance models and actual behaviors significantly affects the ability of occupants to escape safely in the case of a fire. The disparity is primarily caused by three factors: fire characteristics, human characteristics and building characteristics.

- These models are based on the assumption that fire growth in a building follows the standard fire curve. In practice, fire growth depends on the materials used in building construction [90]. Thus, the combustion speed may be ultra-fast. Different fire growths produce different evacuation behaviors.
- Response performance models assume that occupants escape immediately upon hearing a fire alarm. These models also assume a constant walking speed for individuals in the course of the evacuation. However, the pre-movement period can exceed the evacuation time. Social rules strongly influence individual reaction times. While escaping from an incident, individuals can be confronted with many emergencies and therefore walk more slowly than they would under normal conditions. Thus, the occupants' walking speed is not consistent during evacuation [83, 91, 92].
- Designers of green escape route signs and researchers developing response performance models assume that occupants follow these signs to escape. While occupants notice the color, the pictogram and the location of the signs, they usually do not follow the signs in the course of escaping [83, 93]. Response performance models are also based on the paradigm that occupants escape via the nearest exit. However, in real-life incidents, people escape via familiar exit routes [94, 95]. Thus, the layout of the building interior significantly influences human behaviors under emergency conditions.

Fire safety policies are based on these response performance models; however, the disparities between these models and actual behaviors cast doubt on the soundness of the basic principles and the subsequent effectiveness of these policies. Therefore, further studies or scientific experiments should be conducted on human behaviors in real incidents to clarify the reciprocal influence between fire characteristics and human characteristics and between building characteristics and human characteristics from the ignition stage to the end of the evacuation process. New sound basic principles of fire safety policy should consequently be developed in the near future from real observed behaviors. In addition, analyzing detailed information on decision-making for evacuation in sufficient depth can stimulate the development of practical models for the fire prevention design of buildings.

4 Conclusions

During a fire, people perceive fire cues to varying degrees. These perceived cues cause individuals to create mental models, which are used to interpret the situation as safe or presenting a fire risk. This interpretation process is affected by several factors, including the features of the fire cues, the architecture of the building in which the fire occurs and personal characteristics. Individuals then engage in decision-making processes to identify protective actions and to create an adaptive plan or strategy. The dynamic quality of fire situations results in a highly complex path from cue perception to response.

Acknowledgment. The research presented here is part of the "The influence of remote interactive mode of dynamic spatial information on path-finding decision-making in high-rise building fire" project, which is supported by the Beijing Natural Science Foundation. The authors would like to acknowledge the support of the Beijing Natural Science Foundation for this project (9172008).

References

1. Kuligowski, E.D.: Modeling human behavior during building fires. NIST Technical Note 1619. National Institute of Standards and Technology, Gaithersburg (2008)
2. Hall, J.R.: How many people can be saved from home fires if given more Time to escape? Fire Technol. **40**, 117–126 (2004)
3. Kobes, M., Post, J., Helsloot, I., Vries, B.: Fire risk of high-rise buildings based on human behavior in fires. In: Conference Proceedings FSHB 2008. First International Conference on fire Safety of High-rise Buildings. Bucharest, Romania, pp. 07–09, May 2008
4. Tavares, R.M., Galea, E.R.: Evacuation modeling analysis within the operational research context: a combined approach for improving enclosure designs. Build. Environ. **44**, 1005–1016 (2009)
5. Tavares, R.M., Tavares, J.M.L., Parry-Jones, S.L.: The use of a mathematical multi-criteria decision-making model for selecting the fire origin. Build. Environ. **43**, 2090–2100 (2008)
6. Sime, J.D.: Design Against Fire: An Introduction to Fire Safety Engineering Design. Stollard, P., Johnston, L. (eds.), pp. 56–87. E and F. N. Spon, London (1994)
7. Proulx, G., Reid, I.: Occupant behavior and evacuation during the Chicago Cook County Administration Building fire. J. Fire. Prot. Eng. **16**(4), 283–309 (2006)

8. Averill, J.D., et al.: Federal building and fire safety investigation of the world trade center disaster: occupant behavior, egress, and emergency communications. NIST NCSTAR, pp. 1–7, NRCC-48362. National Institute of Standards and Technology, Gaithersburg (2005). http://wtc.nist.gov/oct05NCSTAR1-7index.htm
9. Fahy, R.F., Proulx, G.: Human behavior in the world trade center evacuation. In: Proceeding of the Fifth International Symposium on Fire Safety Science, pp. 713–724 (1997)
10. Slovic, P.: Trust, emotion, sex, politics, and science: surveying the risk-assessment battlefield. Risk Anal. **19**(4), 689–701 (1999)
11. Ballard, G.M.: Industrial risk: safety by design. In: Ansell, J., Wharton, F. (eds.) Risk: Analysis, Assessment and Management, pp. 95–104. Wiley, Chichester (1992)
12. Douglas, M., Wildavsky, A.: Risk and Culture: An Essay on the Selection of Technical and Environmental Dangers. University of California Press, Berkley (1982)
13. Short, J.F.: The social fabric of risk: towards the social transformation of risk analysis. Am. Soc. Rev. **42**(6), 711–725 (1984)
14. Sjöberg, L., Moen, B.E., Rundmo, T.: Explaining Risk Perception: An Evaluation of the Psychometric Paradigm. Rotunde Publicationer, Trondheim (2004)
15. Rosa, E.A.: The logical structure of the social amplification of risk framework (SARF): metatheoretical foundation and policy implications. In: Pidgeon, N., Kasperson, R.E., Slovic, P. (ed.) The Social Amplification of Risk, pp. 47–79. Cambridge University Press, Cambridge (2003)
16. Communities and Local Government. Understanding people's attitudes towards fire risk–final report to communities and local government. Fire Research Series 13/2008. Department of Communities and Local Government, London
17. Spangler, M.B.: Policy issues related to worst case risk analyses and the establishment of acceptable standards of de minimis risk. In: Covello, V.T., Lave, L.B., Moghissi, A., Uppuluri, V.R.R. (eds.) Uncertainty in Risk Assessment, Risk Management, and Decision Making, pp. 1–26. Plenum Press, New York (1984)
18. Garvin, T.: Analytical paradigms: the epistemological distances between scientists, policy makers, and the public. Risk Anal. **21**(3), 443–455 (2001)
19. Adams, J.: Risk. UCL Press, London (1995)
20. Slovic, P.: The Perception of Risk. Earthscan Publications Ltd., London (2000)
21. Tillander, K.: Utilization of statistics to assess fire risk in buildings. Dissertation for the degree of Doctor of Science in Technology at Helsinki University. VTT Building and Transport, Espoo, Finland (2004)
22. Aguirre, B.E.: Emergency evacuations, panic, and social psychology: commentary on understanding mass panic and other collective responses to threat and disaster. Article #402. University of Delaware, Disaster Research Center, Newark (2005)
23. Brennan, P.: Modeling cue recognition and pre-evacuation response. In: Proceedings – 6th International Symposium of Fire Safety Science, pp. 1029–1040. International Association for Fire Safety Science, London (1999)
24. Kuligowski, E.D.: The process of human behavior in fires. NIST Technical Note 1632. National Institute of Standards and Technology, Gaithersburg (2009)
25. Blanchard-Boehm, R.D.: Understanding public response to increased risk from natural hazards: application of the hazards risk communication framework. Int. J. Mass Emerg. Disasters **16**, 247–278 (1998)
26. Klein, G.: Sources of Power: How People Make Decisions. The MIT Press, Cambridge (1999)
27. Bruck, D.: The who, what, where and why of waking to fire alarms: a review. Fire Saf. J. **36**, 623–639 (2001)

28. Bruck, D., Thomas, I.: Waking Effectiveness of Alarms (Auditory, Visual and Tactile) for Adults Who are Hard of Hearing. The Fire Protection Research Foundation, Quincy (2007)
29. Proulx, G., Reid, I., Cavan, N.R.: "Human Behavior Study" Cook County Administration Building Fire, Chicago, IL. National Research Council of Canada, Ottawa (2003)
30. Ozel, F.: The role of time pressure and stress on the decision process during fire emergencies. In: Proceedings of the First International Symposium on Human Behavior in Fire, London, England, pp. 191–200 (1998)
31. Jin, T.: Studies on human behavior and tenability in fire smoke. In: Hasemi, Y. (ed.) Proceedings of the Fifth International Symposium, London, England, pp. 3–21 (1997)
32. Mileti, D.S., Darlington, J.D.: Societal response to revised earthquake probabilities in the San Francisco bay area. Int. J. Mass Emerg. Disasters **13**, 119–145 (1995)
33. Kobes, M., Helsloot, I., De Vries, B., Post, J.G.: Building safety and human behavior in fire: a literature review. Fire Saf. J. **45**, 1–11 (2010)
34. Sime, J.D.: An occupant response shelter escape time (ORSET) model. Saf. Sci. **38**, 109–125 (2001)
35. Mileti, D.S., Fitzpatrick, C.: Causal sequence of risk communication in the park field earthquake prediction experiment. Risk Anal. **12**(3), 393–400 (1992)
36. Lindell, M.K., Perry, R.W.: Communicating Environmental Risk in Multiethnic Communities. Thousand Oaks (2004)
37. Mileti, D.S., Darlington, J.D.: The role of searching in shaping reactions to earthquake risk information. Soc. Prob. **44**(1), 89–103 (1997)
38. Weick, K.E.: Sensemaking in Organizations. Sage Publications, Thousand Oaks (1995)
39. Weick, K.E.: The collapse of sensemaking in organizations: the Mann Gulch disaster. Adm. Sci. Q. **38**, 628–652 (1993)
40. Rudolph, J.W., Morrison, J.B.: Confidence, error, and ingenuity in diagnostic problem solving: clarifying the role of exploration and exploitation. In: Proceedings of the Academy of Management Annual Meeting (2007)
41. Rudolph, J.W., Raemer, D.B.: Diagnostic problem solving during simulated crises in the OR. Anesth. Analg. **98**, 34 (2004)
42. Deloitte. From risk perception to safe behavior. Article on the Safety Institute of Australia Ltd. (2006). http://www.sia.org.au/downloads/SIGs/Resources/From_Risk_Perception_to_Safe_Behaviour.pdf
43. O'Connor, D.J.: Integrating human behavior factors into design. In: Fire Protection Engineering, pp. 8–20 (2005)
44. Proulx, G.: Occupant behavior and evacuation. In: Proceedings of the 9th International Fire Protection Symposium, Munich, pp. 219–232 (2001)
45. Liu, T., Jiao, H.: How does information affect fire risk reduction behaviors? Mediating effects of cognitive processes and subjective knowledge. Nat. Hazards **90**(3), 1461–1483 (2018)
46. Proulx, G.: Playing with fire: understanding human behavior in burning buildings. ASHRAE J. **45**(7), 33–35 (2003)
47. Canter, D., Donald, I., Chalk, J.: Pedestrian behavior during emergencies underground: the psychology of crowd control under life threatening circumstances. In: Vardy, A. (ed.) Safety in Road and Rail Tunnels, Bedford, pp. 135–150 (1992)
48. Phil-Sik, J.: Designing acoustic and non-acoustic parameters of synthesized speech warnings to control perceived urgency. Int. J. Ind. Ergon. **37**(3), 213–223 (2007)
49. Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G.: "Risk as analysis and risk as feelings: some thoughts about affect", reason, risk and rationality. Risk Anal. **24**(2), 311–322 (2004)

50. Kuligowski, E.D.: Terror defeated: occupant sense-making, decision-making and protective action in the 2001 World Trade Center disaster. Ph.D. dissertation. University of Colorado at Boulder, Boulder (2011)
51. Proulx, G.: Occupant responses to fire alarm signals. NFPA 72 Fire Alarm Code, Supplement 4. Quincy, MA (2000)
52. Tong, D., Canter, D.: The decision to evacuate: a study of the motivations which contribute to evacuation in the event of fire. Fire Saf. J. **9**, 257–265 (1985)
53. Mileti, D.S., O'Brien, P.W.: Public response to aftershock warnings. Geological Survey. U.S. Department of the Interior, Washington, D.C (1993)
54. Proulx, G.: As of 2000, "What do we know about occupant behavior in fire?" In: Proceeding of the Technical Basis for Performance Based Fire Regulations, United Engineering Foundation Conference, pp. 127–129 (2001)
55. Gioia, D.A., Poole, P.P.: Scripts in organizational behavior. Acad. Manag. Rev. **9**, 449–459 (1984)
56. Burns, K.: Mental models and normal errors. In: Montgomery, H., Lipshitz, R., Brehmer, B. (eds.) How Professionals Make Decisions, pp. 15–28. Erlbaum Associates, Mahwah (2005)
57. Mileti, D.S., Sorensen, J.H.: Communication of emergency public warnings: a social science perspective and state-of-the-art assessment. Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge (1990)
58. Gigerenzer, G., Selten, R.: Bounded Rationality: The Adaptive Toolbox. The MIT Press, Cambridge (2001)
59. Gwynne, S., Galea, E.R., Lawrence, P.J., Filippidis, L.: Modeling occupant interaction with fire conditions using the building EXODUS evacuation model. Fire Saf. J. **36**, 327–357 (2001)
60. Zakay, D.: The impact of time perception processes on decision making under time stress. In: Svenson, O., John Maule, A. (eds.) Time Pressure and Stress in Human Judgment and Decision Making, pp. 59–72. Plenum Press, New York (1993)
61. Karau, S.J., Kelly, J.R.: The effects of time scarcity and time abundance on group performance quality and interaction process. J. Exp. Soc. Psychol. **28**, 542–571 (1992)
62. Vaughan, D.: The dark side of organizations: mistake, misconduct, and disaster. Ann. Rev. Sociol. **25**, 271–305 (1999)
63. Orasanu, J., Fischer, U.: Finding decisions in natural environments: the view from the cockpit. In: Zsambok, C., Klein, G. (eds.) Naturalistic Decision Making, pp. 343–358. Erlbaum, Mahwah (1997)
64. Janis, I.L., Mann, L.: Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment. Free Press, New York (1977)
65. Kuligowski, E.D., Omori, H.: General Guidance on Emergency Communication Strategies for Buildings - 2nd Edition NIST TN – 1827. US Department of Commerce, National Institute of Standards and Technology (2014)
66. Kuligowski, E.: Predicting human behavior during fires. Fire Technol. **49**(1), 101–120 (2013)
67. Reneke, P.A.: NIST IR 7914 Evacuation Decision Model. NIST Interagency/Internal Report. US Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MS, USA (2013)
68. Siebeneck, L.K., Cova, T.J.: Spatial and temporal variation in evacuee risk perception throughout the evacuation and return-entry process. Risk Anal.: Int. J. **32**(9), 1468–1480 (2012)
69. Woody, E.Z., Szechtman, H.: A biological security motivation system for potential threats: are there implications for policy-making? Front. Hum. Neurosci. **7**, 1–5 (2013)

70. Hinds, A.L., et al.: The psychology of potential threat: properties of the security motivation system. Biol. Psychol. **85**(2), 331–337 (2010)
71. Arru, M., Mayag, B., Negre, E.: Early-warning system perception: a study on fire safety. In: Proceedings of the ISCRAM 2016 Conference Intelligent Decision Support in the Networked Society, Rio de Janeiro, Brazil (2016)
72. Comes, T., Mayag, B., Negre, E.: Beyond early: decision support for improved typhoon warning systems. In: Proceedings of the ISCRAM 2015 Conference, Kristiansand, Norway, vol. 12 (2015)
73. Kuligowski, E.D., Hoskins, B.L.: Occupant behavior in a high-rise office building fire. NIST Technical Note 1664. National Institute of Standards and Technology, Gaithersburg (2010)
74. Purser, D.A., Bensilum, M.: Quantification of behavior for engineering design standards and escape time calculations. Saf. Sci. **38**, 157–182 (2001)
75. Bryan, J.L.: Behavioral response to fire and smoke. In: DiNenno, P.J. (ed.) The SFPE Handbook of Fire Protection Engineering, 3rd edn, vol. 3, pp. 320–354. National Fire Protection Association (2008)
76. Canter, D., Breaux, J., Sime, J.: Domestic, multiple occupancy and hospital fires. In: Canter, D. (ed.) Fires and Human Behaviour, pp. 117–136. Wiley, New York (1980)
77. Proulx, G.: Evacuation time and movement in apartment buildings. Fire Saf. J. **24**, 229–246 (1995)
78. O'Brien, P.W., Atchison, P.: Gender differentiation and after shock warning response. In: Enarson, E., Morrow, B.H. (eds.) The Gendered Terrain of Disaster: Through Women's Eyes, Westport, CT, pp. 173–180 (1998)
79. Enarson, E., Scanlon, J.: Gender patterns in flood evacuation: a case study of couples in Canada's Red River Valley. Appl. Behav. Sci. Rev. **7**(2), 103–124 (1999)
80. Bryan, J.L.: Behavioral response to fire and smoke. In: DiNenno, P.J. (ed.) The SFPE Handbook of Fire Protection Engineering, Quincy, MA, 3rd edn, vol. 3, pp. 315–341 (2002)
81. Purser, D.A., Gwynne, S.M.V.: Identifying critical evacuation factors and the application of egress models. In: Franks, C.A., Grayson, S. (eds.) Proceedings of the 11th International Inter-flam Conference, pp. 203–214. Inter-science Communications Ltd., London (2007)
82. Lindell, M.K., Lu, J.C., Prater, C.S.: Household decision making and evacuation in response to Hurricane Lili. Nat. Hazards Rev. **6**(4), 171–179 (2005)
83. Mileti, D.S., O'Brien, P.W.: Warnings during disasters: normalizing communicated risk. Soc. Prob. **39**(1), 40–57 (1992)
84. Isobe, M., Helbing, D., Nagatani, T.: Many-particle simulation of the evacuation process from a room without visibility. Phys. Rev. 69 (2004)
85. Nagai, R., Nagatani, T., Isobe, M., Adachi, T.: Effect of exit configuration on evacuation of a room without visibility. Physica **343**, 712–724 (2004)
86. Nilsson, D., Johnansson, A.: Social influence during the initial phase of a fire evacuation – analysis of evacuation experiments in a cinema theatre. Fire Saf. J. **44**(1), 71–79 (2009)
87. Johnson, C.W.: Lessons from the evacuation of the world trade center, 9/11 2001 for the development of computer-based simulations. Cogn. Technol. Work **7**, 214–240 (2005)
88. Musse, S.R., Thalmann, D.: A model of human crowd behavior: group inter-relationship and collision detection analysis. In: Thalmann, D., van de Panne, M. (eds.) Computer Animation and Simulation. EUROGRAPH, pp. 39–51. Springer, Vienna (1997). https://doi.org/10.1007/978-3-7091-6874-5_3
89. Le Bon, G.: Psychologies des Foules. PUF, Paris (1971)
90. Zarboutis, N., Marmaras, N.: Investigating crowd behavior during emergency evacuations using agent-based modeling. In: Proceedings of the 24th European Annual Conference on Human Decision Making and Manual Control, Athens, pp. 1345–1357 (2005)

91. Chang, C.H., Huang, H.C.: A water requirements estimation model for fire suppression: a study based on integrated uncertainty analysis. Fire Technol. **41**, 5–24 (2005)
92. Bryan, J.L.: A selected historical review of human behavior in fire. J. Fire. Prot. Eng. **16**, 4–10 (2002)
93. Sime, J.D.: Crowd psychology and engineering. Saf. Sci. **21**, 1–14 (1995)
94. Ouellette, M.J.: Visibility of exit signs. Prog. Archit. **74**(7), 39–42 (1993)
95. Graham, T.L., Roberts, D.J.: Qualitative overview of some important factors affecting the egress of people in hotel fires. Hosp. Manag. **19**, 79–87 (2000)
96. Benthorn, L., Frantzich, H.: Fire alarm in a public building: how do people evaluate information and choose evacuation exit? Department of Fire Safety Engineering, Lund Institute of Technology, Lund University, Sweden (1996)
97. Sullivan-Wiley, K.A., Short Gianotti, A.G.: Risk perception in a multi-hazard environment. World Dev. (2017). https://www.sciencedirect.com/science/article/abs/pii/S0305750X17301195

Music Valence and Genre Influence Group Creativity

Hosseini Sarinasadat[1], Yuki Hattori[1], Yoshihiro Miyake[1],
and Takayuki Nozawa[2](✉)

[1] Tokyo Institute of Technology, Yokohama, Japan
[2] Tokyo Institute of Technology, Tokyo, Japan
nozawa.t.ac@m.titech.ac.jp

Abstract. Previous studies on the effects of music listening on cognitive and affective functioning has focused on individuals, and the influence of back ground music (BMG) on group creativity is yet to be explored. Here, based on the meditative impact of music on emotion and communication, we hypothesized the possible influence of BGM on indices of group creativity, and specifically investigate the effects of two factors of music, valence and genre, on cooperation and divergent creativity. In our study, 15 pairs of participants conducted Alternative Uses Task (AUT)-based communication while listening to four combination of instrumental music tracks or no music. To probe into the mechanism of group creativity enhancement, we assessed the interbrain synchrony using functional near-infrared spectroscopy (fNIRS)-based hyperscanning, and the non-verbal communication by using accelerometers to measure head movements synchronization (HMS) of dyads. Our results suggested that positive valence music enhance cooperation, while upbeat genre enhance cooperation to converge the ideas.

Keywords: Applied cognitive psychology · Team working · Communication · Creativity · Physical synchronization

1 Introduction

Music is what people are willing to spend their time on and there are different reasons behind it, from relieving tension to passing the time [1] or even controlling the moods [2]. Although people might often prefer to sit down and listen to music deliberately, Renfrow and Gosling reported a wide variety of activities when people might listen to music [3]. As an important aspect of human daily life, many studies have been conducted on the role of music on a wide range of individual behavior, informing us with the positive influence of music on the sense of helpfulness, task involvement encouragement and coping with perceived stress [4]. Studies reported effect of specific music on individual spatial abilities [5, 6], though controversy on the source and reproducibility of the effect makes it difficult to come up with a single conclusion [7]. A recent study [8] showed that listening to "happy music" enhance divergent thinking. Similarly, Ilie *et al.* examined the cognitive changes in term of creativity and mentioned over the effect of music type on individual creativity [9]. Although these results

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 410–422, 2019.
https://doi.org/10.1007/978-3-030-22507-0_32

might explain music impact on some levels of individual creative cognition, the field of the possible effects of music on group communication is to be investigated further.

Cross and Morley argued over the music capacity to sustain social interactions [10]. Performance of joint music making on cooperation has been studied before [11]. Brown *et al.* [12] investigated how music impact on cooperation. In that article they mentioned the use of music isometric rhythms to enhance group synchronization [12]. Also, Lang *et al.* [14], based on the observations that interpersonal coordination results in subsequent social bonding enhancement [13], referred to rhythm impact on group coordination enhancement as a route to facilitate positive social behavior and bonding [14]. In complementary experiments, Au *et al.* argued that participant who listened to pleasant music tended to be more confident over the ones who listened to unpleasant music [15] and Greitemeyer presented the music effects on mood and decision behavior [16]. Physiological synchrony is intertwined with emotional rapport [17]. A review article [18] mentioned how music makes brainstem neurons fire synchronously with tempo, and synchronized activities like music encourage social connection. Recently, Bernardi *et al.* [19] showed that listening to simple rhythms makes individuals synchronized in terms of their physiological rhythms, which may lead to rapport and mutual understanding. These findings bring us to the possible influence of music's synchrony-inducing (i.e. homogenizing) effects on group cognition, but would music facilitate or hinder group creativity, and how such an influence can depend on the different types of music?

Group creativity, being the way to enhance the creative productivity through communication, has been an interesting study topic for many years. However, the benefit of group communication on creativity has been on the controversy in many of them. Being in a group would hinder creativity performance in terms of the productivity loss in idea-generating [20]. Perceiving others efforts to be sufficient, apprehensive behavior toward other members judgment [21] and the fact that only one person can talk at a time [22] are the reasons behind this idea. Although being a well-perceived fact, it neglects the productivity impacts of being in a group on creativity [23] with benefits like minimizing member's motivation, energy, and talent losses, and misuse of time [24]. While traditional thoughts believe divergent perspectives increase and homogeneous perspectives decrease group creativity [25], recently there is a hiatus on this view since the convergent process is needed for distinguishing new ideas and unifying them. The new ideas on group creativity explain how groups might benefit from cooperation. Where the individual ability of group members to produce new ideas is important [26] and diversity between members can enhance the comparison between group members to enhance divergent thinking [27]. Group creativity, on the other hand, benefits from its members' cooperation to integrate original perspectives [28]. Use of group to enhance creative productivity can be practical as long as the diversity between members and their shared ideas would not disservice the cooperation level by being in opposition [29].

As noted above, some music types can enhance cooperation between members of a group in general. However, its impacts on group creativity are yet to be investigated. Group creativity, as discussed earlier, is based on the level of individual creativity along with the cooperation tendency of its members. While music effect on group creativity has been scarcely explored so far, previous studies stated that members with high extraversion and sociability traits will experience less level of anxiety during experiments [30] which might further result in convergent thinking facilitation [31]. In their study, they presented a higher level of tendency to be fixed on creativity in high preference interactive participants compared to low preference ones but no trend to produce more unique ideas.

Creative thinking is the combination of producing as much as possible ideas (fluency), in many categories (flexibility), while the ideas remain unique and novel (originality) and needs both of the cooperation and individual creativity of its members to be enhanced. In this article, we aimed to compare the creative performance of interactive groups while listening to different types of music, being positive vs. negative in terms of valence and reflective vs. upbeat in terms of the track's types. Our purpose is to test the impact of different music types on the different group creativity indices and investigate over the music types which would help or hinder individual creativity, and cooperation level of members, resulting in the total group creativity changes.

To probe the processes underlying the effects of music on group creativity, we investigated interbrain synchrony and non-verbal communication (NVC) expressed in the physical interpersonal coordination. In the exploration of brain functioning in experimental and daily human interactions, the hyperscanning technique allows brain activity measurement of two or more people simultaneously [32]. Functional near infrared spectroscopy (fNIRS) allows brain activity measurement during natural communications, with high ecological validity, portability, and cost effectiveness. fNIRS-based hyperscanning studies [33, 34] indicated that the level of cooperation during tasks is correlated to interbrain synchrony in the pre-frontal cortex (PFC), which has been associated with cognitive processes such as working memory and executive function [35]. In this study, we measured the medial and left lateral part of the PFC. In previous studies, the left inferior PFC showed association with the memory process functions during communication [36], while anterior PFC (aPFC) involved in integrating different operations into behavioral goal [37]. Moreover, NVC is the way to make communication without transmission of the words [38]. In this study, to test the effectiveness of music to enhance NVC, we evaluated the head movement synchrony (HMS) [39].

2 Method

2.1 Participants

Thirty international students of Tokyo Institute of Technology including 18 females and 12 males being recruited via flyers and took part in our experiment. All participants being right-handed with normal or corrected-to-normal vision. Participants were grouped into dyads and been called to participate in the experiment based on their answers to an online preparation phase questionnaire, which will be further explained in the next subsection. The study procedure was approved by the Ethics Committees of Tokyo Institute of Technology. All participants were briefed about the experimental procedure and gave written informed consent. They were paid 3000 Yen for their time and effort.

2.2 Selection of Music Stimuli and Dyad Construction

Music asserts its effects through influencing emotions [40]. On the concept of emotion within music, Schimmack and Grob focused on two elements of music: valence (pleasant vs unpleasant) and arousal [41]. In this study, to select a list of music pieces as stimulus, we first fixed our factors of interest on valence and genre. Here, the genre was whether a piece is “reflective” or “upbeat”. The reflective list consisted of classical pieces mostly from famous composers, while upbeat tracks were defined by either country, sound track, and pop music categories [42].

In order to delineate the effects of music valence and genres from other possible confounding factors, we tried to control the degree of familiarity, tempo, and likability of the music pieces as follows. A study [43] named familiarity as a factor to engage listeners on the music and addressed over its impact on involved emotions. As the level of familiarity and emotional perception between members in the groups might be different along with different provoked memories, we decided to use only unfamiliar pieces in our experiment. In addition, several articles [44, 45] counted music tempo as a factor to evaluate the emotional connections of music. The connections between music tempo and physical movements has been also reported [46]. Therefore, we chose music tracks within the range of [95–105] bmp. The judgment of liking a music is through the level of emotion of pleasure. Likable pieces can activate specified parts of the brain regions [43]. Perceived enjoyment depends on individual preference and also related with factors such as familiarity, personality [47]. Therefore, we fixed the likability level between group members on the highest level of chosen music pieces for the experiment setting.

In a preparation phase before the experiment, each participant listened to the first fifteen seconds of one hundred music pieces for candidate stimuli, all instrumental version, and rated their familiarity, likability and perceived mood (valence). The ratings of perceived mood were used to confirm the validity of our identification of valence based on the pitch and key movements of each piece. The pieces were selected by the experimenter to make them equally distributed over the combinations of categories: positive vs. negative valence × reflective vs. upbeat tracks [42]. All of the tracks were put in randomized order on an online survey. Participants who rated at least one same track as low on familiarity, high in likability and very low/very high on mood (for each of negative/positive valence tracks) for each of reflective and upbeat genre were further grouped into dyads and participated in the experiment.

2.3 Task Procedure

Before the experiment, an explanation was given and a practice session has been done. Each experiment had two sessions with three trials in the first and two trails in the second session, with 6 min for each trial and a 10-min break between the sessions. During each trial, following an audio cue, a name of a familiar object (One meter of cotton rope- An egg- A plank of wood- A tennis ball, and A pair of socks) has been shown on a TV screen while one of the four types of tracks (i.e. positive-reflective, positive-upbeat, negative-reflective, negative-upbeat), which were selected in the preparation phase as explained above, or no music, was played as background music. The order of objects and music types were randomized over dyads. Seeing the name of the object, participants started alternative uses tasks (AUT) cooperatively through communication, saying their answers loud enough to be recorded by a voice recorder. After each trial, another audio cue was presented and participants answered a questionnaire about their mood. Also, it should be noted that these chosen objects were assumed to be not different in difficulty in thinking and discussing alternative uses based on the results of pilot experiments which was confirmed after assessment of indices of creativity on our main experiments.

2.4 Evaluation of Creativity Indices

To evaluate creativity, ideas captured during experiments via voice recorder were coded. We further assessed the indices of creativity in four different categories of fluency, originality, flexibility and index of convergence (IOC) [31]. The total number of ideas mentioned during each trial has been calculated as the group fluency. The originality was assessed by the average of repetition likelihood of each idea within all groups [48]. Based on the category identification of the generated ideas, flexibility was defined as the total number of the visited categories during each trail. The IOC was a measure of cooperation behavior of dyads and calculated by dividing the number of times each dyad stayed on the same category over the times they deferred the category or totally moved to a new category.

2.5 Characterization of Communication with Inter Brain Synchronization

Dyads in each group wore a portable functional near-infrared spectroscopy (fNIRS) device (HOT-1000; Hitachi Hitech, Co.) to measure their brain activities during each trial. Participants have been instructed to avoid unnecessary movements as much as possible, to reduce the possibility of unwanted noises. According to the international 10–20 system, we placed the center (i.e. channel) of the two optodes of fNIRS device on FP2 (left channel) and FPz (medial channel), respectively for both of the dyads, based on their head shape. The device sampled changes in the absorption of near-infrared light at a sampling rate of 10 Hz. The data were preprocessed in order to reduce the effect of noise. For each set of data from participants, after linear detrending to remove the trend, we applied Savitzky-Golay smoothing filters with an order of 3 and framelen of 41 following to band pass Gaussian filter. In the last step to assess the inter-brain synchronization of each dyad, we performed the wavelet transform coherence (WTC) in the timescale of [10 to 100 s] on the filtered data of each dyad.

2.6 Characterization of Communication with Head Movement Synchronization

To assess the data of head movement, a small accelerometer (TSND121; ATR-Promotions) was attached to the fNIRS device, at the position of FPz. We set the sampling time at 10 ms. After taking raw data of head movement of each participants from the attached accelerometers, by applying Spearman's rank correlation we calculated the head motion time lag between dyads in each group. Doing this we identified their level of head movement synchronization (HMS) [39] and assessed the head nodding data of each group during each trial. We further separated the signal into two frequency ranges of low [1–1.5 Hz] and high [3.5–5 Hz] [49] and analyzed respectively.

3 Results

3.1 Effects of Music Types on Creativity Indices

One-way repeated measures ANOVA for the five music conditions showed significant difference over conditions in terms of fluency (F(4,56) = 22.48, p < 0.001), originality (F(4,56) = 14.39, p < 0.001), and IOC (F(4,56) = 4.00, p = 0.004) but not for flexibility (F(4,56) = 0.14, p = 0.96).

Excluding the data of no-music condition, two-way repeated measures ANOVA on fluency with factors of genre and valence revealed significant main effects in both of the valence (F(1,14) = 38.98, p < 0.001) and genre (F(1,14) = 7.55, p = 0.016). There was significant interaction between valence and genre (F(1,14) = 19.37, p = 0.001). Also, as for the originality score, significant main effects in both of the valence (F(1,14) = 10.25, p = 0.006) and genre (F(1,14) = 29.34, p < 0.001) was observed. There was a significant interaction between the two factors (F(1,14) = 19.05, p = 0.001). On the other hand, two-way repeated measures ANOVA on IOC score revealed significant main effect of genre (F(1,14) = 9.42, p = 0.008), with upbeat genre tracks leading to higher IOC score. There was no significant interaction between the valence and genre or main effect of the valence for IOC.

Observing these results, separate pairwise t-tests were done between positive-upbeat music vs no music control condition, showing significant enhancing effects in fluency (t(14) = 13.93, p < 0.001) originality (t(14) = 6.75, p < 0.001) and IOC (t(14) = 2.06, p = 0.025). Summary statistics of the creativity performance indices are illustrated on Table 1.

Table 1. Indices of creativity task performance

Measures	No music	Reflective genre		Upbeat genre	
		Negative valence	Positive valence	Negative valence	Positive valence
Fluency					
M	−0.87	−0.26	−0.03	−0.37	1.48
SD	0.47	0.71	0.88	0.66	0.45
Originality					
M	−0.40	−0.33	−0.63	−0.36	1.17
SD	0.70	0.64	0.70	0.89	0.48
IOC					
M	0.064	−0.63	−0.19	0.21	0.79
SD	0.76	0.76	0.77	1.07	1.14
Flexibility					
M	0.06	0.002	0.01	−0.14	−0.08
SD	0.89	0.80	0.94	0.99	1.14

3.2 Effects of Music Types on Nonverbal Communication Measures

Non-verbal communication was evaluated by calculating HMS during each trial for each dyad. To test music effect on physical synchrony, one-way repeated measures ANOVA for the five conditions was conducted. There was a significant difference over the conditions (F(4,56) = 2.64, p = 0.043; Fig. 1). Also, two-way repeated measures ANOVA with genre and valence as factors of effect has been done. Quasi-significant effect was observed for valence (F(1,14) = 3.33, p = 0.089), but not for the other factor or their interaction.

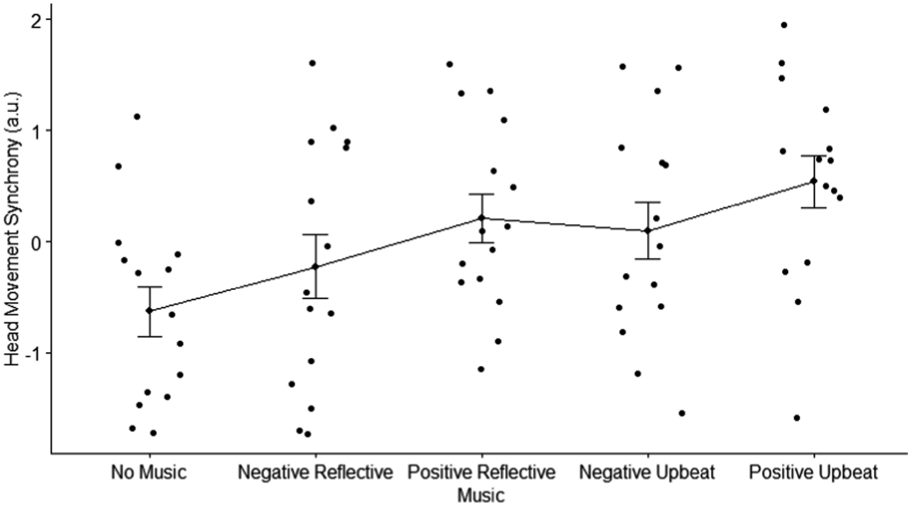


Fig. 1. Relation between back ground music types and head movement synchrony (HMS). Data illustrated a significant enhancement of HMS for all music combinations (F(4,56) = 2.64, p = 0.043). Error bars show standard error of the mean.

To investigate whether the difference was more significant in low frequency ranges or higher frequency ranges, separate one-way ANOVAs have been conducted. The results indicated the significant effect of music over HMS only in the low frequency range (F(4,56) = 3.19, p = 0.020).

As of the inter brain synchronization the results of ANOVA on IBS showed no significant difference between five conditions on either of the medial (F(4,56) = 0.41, p = 0.80) or the left lateral channel (F(4,56) = 1.08, p = 0.37). Summary statistics of the inter brain synchrony are shown on Table 2.

Table 2. Effects of music types on IBS

Measures	No music	Reflective genre		Upbeat genre	
		Negative valence	Positive valence	Negative valence	Positive valence
Medial channel					
M	−0.01	0.11	−0.12	−0.08	0.01
SD	0.58	0.43	0.48	0.48	0.45
Left lateral channel					
M	−0.02	0.38	0.06	−0.15	0.21
SD	0.43	0.46	0.38	0.43	0.53

4 Discussion

This study has addressed effect of background music on the group creativity. We hypothesized that the group creativity process might be strongly affected by different combinations of music features. In the experiment, we investigated whether listening to a specific type of music as compared to no music control condition might enhance group creativity. To test this hypothesis, we manipulated four types of music that varied on two terms of genre (reflective vs upbeat) and valence (positive vs negative), while controlling tempo, familiarity, and likability by each dyad. This control was because, in a previous study [9] with manipulations of rate, pitch height, and intensity of music they stated the main effect of rate over perceived arousal and pitch and intensity as the experiential quality of emotion over valence.

In the main part of our result, there is an evidence that background music listening in all four combinations, as compared to control condition facilitated the group creativity in term of fluency but not the other three indices of interest. To explain over this; literature on the music has proven the effect of music listening to alter mood [50, 51], and on the studies of group creativity, behaviors such as negative attitudes, judgments and evaluative behavior during discussion has been addressed as disadvantages [52, 53]. On the other side of explanation, articles on the mood suggest that positive mood enhance the generation of the ideas [54]. These in combination might suggest that music might have decreased the judgement and stress level during group creativity task with mood manipulation and thus resulted into generation of more ideas, though all music types might not have positive influence on originality, convergence index, or flexibility of the ideas.

Our next questions were what effects music valence and genre might have on group creativity, and whether the effects would be the same or not for the different creativity indices. Previous researches on the individual divergent creativity suggested the positive effect of positive valence music. When participants performed the AUT task as a team in the positive valence conditions, the total amount of shared idea has increased compared to negative valence conditions, but this effect was not apparent in the cases of originality or IOC indices for the reflective positive valence pieces. While there was no difference between conditions for flexibility, as for the IOC the main effect of the genre was statistically significant not only compared to reflective pieces but also on no music condition. IOC is a more detailed group creativity analysis to investigate the convergent tendencies of groups [31] and results indicate the advantage of listening to upbeat background tracks on convergent thinking.

Positive-upbeat music improved group creativity in all indices expect for flexibility. While to our knowledge there are only evidence of positive music enhancing individual creativity and no article on the effect of positive and upbeat music on group creativity, an interpretation can be provided through the possible enhancing effect of such music on cooperation. A previous study [55] supported that cooperation between participants resulted into a higher level of originality in their creativity task. Therefore, positive-upbeat music might have enhanced cooperation into especially higher level, leading to the general performance enhancement. This hypothesis can be further supported by the evidence of correlation between extraversion and sociability traits with upbeat music [46].

In the other part of our results, we show the higher enhancing effect of all music combinations compared to the no music one on the head movement synchrony (HMS). Our hypothesis of music enhancing non-verbal communication has been supported by this result. Also, a certain trend toward significance effect of positive valence tracks on head movement synchrony along with its positive impact on creativity task fluency index, might suggest the effect of positive valence music on cooperation between dyads, which mediated higher group creativity.

The pre-frontal cortex (PFC) has been associated with social cognition, decision-making, and goal-maintenance in several articles [56]. However, our results showed no significant effect of music on inter-brain synchrony at any of the two tested channels.

Limitation of The Current Study And Future Research. Although there are more people who might benefit from group creativity activities, our participants are limited to international graduate students with the age range from 24 to 32 years old who are at a highly educated level. This specific nature of the sample, with the possibility of language barrier (majority of participants were not native English speakers) and cultural difference between participants (majority of the dyads were consisted of different ethnical participants), in addition to the limited sample size, makes it difficult to infer the generality of the observed results. We expect that the magnitude of the enhancement of group creativity by music might be different for groups of participants who share same language and are from the same ethnicity. Further study is needed to investigate such possibility. In addition, considering the possible effect of conscious/unconscious process on creativity [57], the effects of background music may also change in less consciously demanding settings than the current group AUT task, for example creativity observed in casual chat.

While in this study we tried to control some factors of music such as familiarity, tempo, and likability, the effect of these factors can bring interesting opportunities for further research. For example, previous studies showed higher impact of familiar songs on individual's mood and cognition [43], so it would be reasonable to assume that familiar music would exert larger influence on group creativity as well. While we used only instrumental music pieces, the music lyrics can be an additional influential factor, which is also connected with familiarity. Furthermore, individual personality traits such as group tendency or habitual involvement in music listening can also modulate the effect of music on group creativity.

Although this study brought us some new evidence on the effect of music on group creativity at the levels of behavioral performance (creativity indices) and nonverbal communication (HMS), we could not observe the contribution of music valence or genre on inter brain synchrony. This could have been caused by smaller coverage of measured brain regions (cf. [55] for example). Another direction yet to be pursued can be to analyze contributions of inter-brain and non-verbal synchronization in combination with music to total group creativity. In our future study, we suggest combining dyadic behavioral variables to increase the chance of the better explanatory model of group creativity.

While in this study we used a questionnaire to measure the effectiveness of music to change in perceived mood and emotion in terms of arouse and pleasure in each trial [58], the limited sample size and an insensitive scale of the questionnaire led to failure in capturing consistent mood changes, so we omitted the results on the questionnaire data from the current report. It is possible that the use of objective mood measurement e.g., heart rate, blood pressure, or skin conductance [59], may bring more accurate results.

5 Conclusion

In conclusion, this study presented the effect of music to influence group creativity. The results brought an evidence of the impact of instrumental music to enhance the total number of generated ideas (fluency) between participants. All music combinations prove higher level of head nodding synchrony compared to the no music condition in our experiments and finally although positive valence have some contribution to participants' engagement to the task and cooperation, upbeat genre music facilitate the index of cooperation to convergent ideas on a significant level. Finally, upbeat genre with positive valence music led participant the highest group creativity, presumably through enhancement of higher engagement, mutual understanding or reciprocal rapport level.

Acknowledgements. This study was partially supported by KAKENHI (Grant Numbers 15H01771 and 17H01753) from JSPS/MEXT, Japan, and by the Center of Innovation Program from Japan Science and Technology Agency (JST), Japan.

References

1. Gantz, W., Gartenberg, H.M., Pearson, M.L., Schiller, S.O.: Gratifications and expectations associated with pop music among adolescents. Pop. Music Soc. **6**, 81–89 (1978)
2. Lonsdale, A.J., North, A.C.: Why do we listen to music? A uses and gratifications analysis. Br. J. Psychol. **102**, 108–134 (2011)
3. Gosling, S.D., Rentfrow, P.J., Swann Jr., W.B.: A very brief measure of the big-five personality domains. J. Res. Pers. **37**, 504–528 (2003)

4. Fried, R., Berkowitz, L.: Music hath charms… and can influence helpfulness 1. J. Appl. Soc. Psychol. **9**, 199–208 (1979)
5. Rauscher, F.H., Shaw, G.L., Ky, K.N.: Music and spatial task performance. Nature **365**, 611 (1993)
6. Rauscher, F., Shaw, G.L., Ky, K.N.: Listening to Mozart enhances spatial-temporal reasoning: towards a neurophysiological basis. Neurosci. Lett. **185**, 44–47 (1995)
7. Waterhouse, L.: Multiple intelligences, the Mozart effect, and emotional intelligence: a critical review. Educ. Psychol. **41**, 207–225 (2006)
8. Ritter, S.M., Ferguson, S.: Happy creativity: listening to happy music facilitates divergent thinking. PLoS ONE **12**, e0182210 (2017)
9. Ilie, G., Thompson, W.F.: Experiential and cognitive changes following seven minutes exposure to music and speech. Music Percept. Interdisc. J. **28**, 247–264 (2011)
10. Cross, I., Morley, I.: The evolution of music: theories, definitions and the nature of the evidence. In: Communicative Musicality: Exploring the Basis of Human Companionship, pp. 61–81 (2009)
11. Dunbar, R.I., Kaskatis, K., MacDonald, I., Barra, V.: Performance of music elevates pain threshold and positive affect: implications for the evolutionary function of music. Evol. Psychol. **10**, 147470491201000403 (2012)
12. Brown, S.: Evolutionary models of music: from sexual selection to group selection. In: Tonneau, F., Thompson, N.S. (eds.) Perspectives in Ethology, vol. 13, pp. 231–281. Springer, Boston (2000). https://doi.org/10.1007/978-1-4615-1221-9_9
13. Hove, M.J., Risen, J.L.: It's all in the timing: interpersonal synchrony increases affiliation. Soc. Cogn. **27**, 949–960 (2009)
14. Lang, M., Shaw, D.J., Reddish, P., Wallot, S., Mitkidis, P., Xygalatas, D.: Lost in the rhythm: effects of rhythm on subsequent interpersonal coordination. Cogn. Sci. **40**, 1797–1815 (2016)
15. Au, K., Chan, F., Wang, D., Vertinsky, I.: Mood in foreign exchange trading: cognitive processes and performance. Organ. Behav. Hum. Decis. Process. **91**, 322–338 (2003)
16. Greitemeyer, T.: Exposure to music with prosocial lyrics reduces aggression: first evidence and test of the underlying mechanism. J. Exp. Soc. Psychol. **47**, 28–36 (2011)
17. Levenson, R.W., Ruef, A.M.: Physiological aspects of emotional knowledge and rapport (1997)
18. Chanda, M.L., Levitin, D.J.: The neurochemistry of music. Trends Cogn. Sci. **17**, 179–193 (2013)
19. Bernardi, N.F., et al.: Increase in synchronization of autonomic rhythms between individuals when listening to music. Front. Physiol. **8**, 785 (2017)
20. Stroebe, W., Diehl, M.: Why groups are less effective than their members: on productivity losses in idea-generating groups. Eur. Rev. Soc. Psychol. **5**, 271–303 (1994)
21. Diehl, M., Stroebe, W.: Productivity loss in brainstorming groups: toward the solution of a riddle. J. Pers. Soc. Psychol. **53**, 497 (1987)
22. Nijstad, B.A., Stroebe, W., Lodewijkx, H.F.: Production blocking and idea generation: does blocking interfere with cognitive processes? J. Exp. Soc. Psychol. **39**, 531–548 (2003)
23. Osborn, A.F.: Applied imagination, principles and procedures of creative thinking (1953)
24. Hackman, J.: The design of work teams. In: Lorsch, J.W. (ed.) Handbook of Organizational (1987)
25. Bantel, K.A., Jackson, S.E.: Top management and innovations in banking: does the composition of the top team make a difference? Strateg. Manag. J. **10**, 107–124 (1989)
26. Taggar, S.: Individual creativity and group ability to utilize individual creative resources: a multilevel model. Acad. Manag. J. **45**, 315–330 (2002)

27. Watson, W.E., Kumar, K., Michaelsen, L.K.: Cultural diversity's impact on interaction process and performance: comparing homogeneous and diverse task groups. Acad. Manag. J. **36**, 590–602 (1993)
28. Harvey, S.: A different perspective: the multiple effects of deep level diversity on group creativity. J. Exp. Soc. Psychol. **49**, 822–832 (2013)
29. Harrison, D.A., Klein, K.J.: What's the difference? Diversity constructs as separation, variety, or disparity in organizations. Acad. Manag. Rev. **32**, 1199–1228 (2007)
30. Camacho, L.M., Paulus, P.B.: The role of social anxiousness in group brainstorming. J. Pers. Soc. Psychol. **68**, 1071 (1995)
31. Larey, T.S., Paulus, P.B.: Group preference and convergent tendencies in small groups: a content analysis of group brainstorming performance. Creativity Res. J. **12**, 175–184 (1999)
32. Montague, P.R., et al.: Hyperscanning: Simultaneous Fmri During Linked Social Interactions (2002)
33. Cui, X., Bryant, D.M., Reiss, A.L.: NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation. NeuroImage **59**, 2430–2437 (2012)
34. Liu, N., Mok, C., Witt, E.E., Pradhan, A.H., Chen, J.E., Reiss, A.L.: NIRS-based hyperscanning reveals inter-brain neural synchronization during cooperative Jenga game with face-to-face communication. Front. Hum. Neurosci. **10**, 82 (2016)
35. Kringelbach, M.L., Rolls, E.T.: The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. Prog. Neurobiol. **72**, 341–372 (2004)
36. Poldrack, R.A., Wagner, A.D., Prull, M.W., Desmond, J.E., Glover, G.H., Gabrieli, J.D.: Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. NeuroImage **10**, 15–35 (1999)
37. Ramnani, N., Owen, A.M.: Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. Nat. Rev. Neurosci. **5**, 184 (2004)
38. LaFrance, M.: Nonverbal synchrony and rapport: analysis by the cross-lag panel technique. Social Psychol. Q. 66–70 (1979)
39. Thepsoonthorn, C., Yokozuka, T., Miura, S., Ogawa, K., Miyake, Y.: Prior knowledge facilitates mutual gaze convergence and head nodding synchrony in face-to-face communication. Sci. Rep. **6**, 38261 (2016)
40. Konecni, V.J.: Does music induce emotion? A theoretical and methodological analysis. Psychol. Aesthet. Creativity Arts **2**, 115 (2008)
41. Schimmack, U., Grob, A.: Dimensional models of core affect: a quantitative comparison by means of structural equation modeling. Eur. J. Pers. **14**, 325–345 (2000)
42. Rentfrow, P.J., Gosling, S.D.: The do re mi's of everyday life: the structure and personality correlates of music preferences. J. Pers. Soc. Psychol. **84**, 1236 (2003)
43. Pereira, C.S., Teixeira, J., Figueiredo, P., Xavier, J., Castro, S.L., Brattico, E.: Music and emotions in the brain: familiarity matters. PLoS ONE **6**, e27241 (2011)
44. Dalla Bella, S., Peretz, I., Rousseau, L., Gosselin, N.: A developmental study of the affective value of tempo and mode in music. Cognition **80**, B1–B10 (2001)
45. Pallesen, K.J et al.: Emotion processing of major, minor, and dissonant chords: a functional magnetic resonance imaging study. Ann. N. Y. Acad. Sci. 1060, 450–453 (2005)
46. Anshel, M.H., Marisi, D.Q.: Effect of music and rhythm on physical performance. Res. Q. Am. Alliance Health Phys. Educ. Recreat. **49**, 109–113 (1978)
47. Juslin, P.N., Liljeström, S., Västfjäll, D., Lundqvist, L.-O.: How does music evoke emotions? Exploring the underlying mechanisms (2010)
48. Dippo, C., Kudrowitz, B.: Evaluating the Alternative Uses Test of Creativity. In: 2013 NCUR (2013)

49. Niewiadomski, R., Mancini, M., Ding, Y., Pelachaud, C., Volpe, G.: Rhythmic body movements of laughter. In: ACM Proceedings of the 16th International Conference on Multimodal Interaction, pp. 299–306 (2014)
50. Balch, W.R., Lewis, B.S.: Music-dependent memory: the roles of tempo change and mood mediation. J. Exp. Psychol. Learn. Mem. Cogn. **22**, 1354 (1996)
51. Gerrards-Hesse, A., Spies, K., Hesse, F.W.: Experimental inductions of emotional states and their effectiveness: a review. Br. J. Psychol. **85**, 55–78 (1994)
52. Sternberg, R.J., Lubart, T.I.: An investment theory of creativity and its development. Hum. Dev. **34**, 1–31 (1991)
53. Guilford, J.P.: Creativity: yesterday, today and tomorrow. J. Creative Behav. **1**, 3–14 (1967)
54. Fredrickson, B.L.: The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. Am. Psychol. **56**, 218 (2001)
55. Xue, H., Lu, K., Hao, N.: Cooperation makes two less-creative individuals turn into a highly-creative pair. NeuroImage **172**, 527–537 (2018)
56. Miller, E.K., Cohen, J.D.: An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. **24**, 167–202 (2001)
57. Ritter, S.M., Van Baaren, R.B., Dijksterhuis, A.: Creativity: the role of unconscious processes in idea generation and idea selection. Think. Skills Creat. **7**, 21–27 (2012)
58. Russell, J.A.: A circumplex model of affect. J. pers. soc. psychol. **39**, 1161 (1980)
59. Salimpoor, V.N., Benovoy, M., Longo, G., Cooperstock, J.R., Zatorre, R.J.: The rewarding aspects of music listening are related to degree of emotional arousal. PLoS ONE **4**, e7487 (2009)

Human Operator Authentication Using Limited Voice Data: A Power Grid Dispatcher Instance

Zheng Wang(✉), Zhen Wang, Yanyu Lu, and Shan Fu

School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
letitbe@sjtu.edu.cn

Abstract. Automatic speaker verification (ASV) has the potential to replace the error-prone and expensive human-based security check to protect the ever increasingly interconnected complex systems, such as power grid system. However, state-of-the-art ASV system relies heavily on a large amount of matched development data and adequate long duration test utterance to maintain the acceptable performance. Unfortunately, such large amounts of data are not always feasible to collect in real world application. In this paper, we propose a new method for i-vector extraction by incorporating historical test information to reduce to requirement of long test utterance duration. The historical tests are weighted by a world MAP estimator and then used in the computation of current test's Baum-Welch statistics. Meanwhile, we modify linear discriminant analysis (LDA) to reduce the requirement of matched development data. In modified LDA training, the variability between development and evaluation data is separated and the objective is to simultaneously minimize the within-class variability and domain variability when maximize the between-class variability. Experiments are conducted on data collected from power grid dispatchers. By adding historical test information, we observe consistent improvement over baseline system especially for shorter duration condition. With modified LDA, at least 63% of performance gap is recovered when system parameters are trained with mismatched development data. Finally, we integrate proposed methods in one system and apply it to power grid dispatching room scenario. Experimental results show our proposed methods achieve fair performance with limited voice data and successfully reduce the amount of data required by ASV system.

Keywords: Speaker verification · Power dispatching · I-vector · Linear discriminant analysis · Short utterance · Domain mismatch

1 Introduction

Power grid is essential for today's society as an enabling infrastructure. The efficiency and safety of power system have major consequences for maintaining stable electricity supply, supporting economic growth and ensuring national security. With the rapid development of technology, a lot of sophisticated automation has been introduced into the power system operation. Since this equipment become more complex and start to

D. Harris (Ed.): HCII 2019, LNAI 11571, pp. 423–437, 2019.
https://doi.org/10.1007/978-3-030-22507-0_33

affect each other, the risk and potential loss of malicious intrusion or attack also increase. Thus, there is an increasing need for verifying the identity of the person regarding authorized to operate the particular machine.

In this situation, conventional human-based authentication such as passwords, tokens, and manual checks is no longer considered to offer high level security alone because human operators are found one of the biggest sources of errors in complex systems [1]. For example, passwords or pin numbers are easily forgotten or forged. And even the most highly trained and alert operators are prone to fatigue and boredom after a long period of continuous work. Therefore, the biometric identification technology can be a useful supplement to existing authentication techniques.

One of the most promising biometric identification technologies is automatic speaker verification (ASV), which is the task of verifying an individual's identity from their voice samples using machine learning algorithms, without any human intervention. Since voice has been one of the most casual means for natural interactions between humans and machines, voice-based systems are easy and intuitive for human operators to use. Further, voice is inherent to individuals and can neither be lost nor stolen which makes it highly accurate and reliable. The availability of low-cost and portable microphones gives it capability of easy integration. ASV has seen significant advancements over the past few decades, giving rise to the successful introduction for various sectors, such as health care, finance and manufacturing industry etc.

Although state-of-the-art i-vector/PLDA based systems exhibit satisfactory performance with adequate speech data [2], a major challenge in ASV is to improve performance with limited voice segments. On the one hand, to achieve fair performance, ASV systems need to be presented with sufficient long utterance (two or three minutes) for enrollment and test i-vectors extraction [3, 4]. Indeed, it is often difficult to acquire such long speech for practice ASV systems because of background noise, voice overlaps or faulty recording devices. Also, there are difficulties related to speaker himself. In fact, unwilling speakers, the state of health, the character of speakers can all contribute to a reduced available amount of speech data. On the other hand, the systems require a large amount of development data to estimate reliable hyper-parameters. Particularly, the success of PLDA modeling depends on the availability of a large set of labeled in-domain data. In most real-life application, collection of such amount of development data from target domain is infeasible. Hence, it is crucial to maintain ASV performance when it is constrained on limited voice data.

Over the years, considerable research effort has been made to overcome such challenges. In [5], the duration variability is mitigated by propagating the posterior covariance of i-vectors to PLDA. However, scoring is computationally expensive in this method. The work in [6] proposed full posterior distribution PLDA to address short duration issue. The work in [7] attempted to improve short utterance system performance by adaptation for i-vector estimation. Also, many techniques are proposed to deal with inadequate target domain data in PLDA modeling. The work in [8] proposes Bayesian adaptation of PLDA models. In [9], unsupervised clustering of i-vectors for adapting covariance matrices of PLDA models is proposed. The work in [10] proposes inter-dataset variability compensation (IDVC) to find a feature space that is more domain independent. In this paper, we propose a new method by incorporating historical test information for short utterance i-vector extraction. In addition, we modify the conventional LDA projection to compensate the domain mismatch before PLDA modeling. In contrast to the existing works address limited utterance length and limited in-domain development data in separate view, we integrate proposed methods in one system and validate it in a real-life power grid dispatching room scenario.

The rest of the paper is organized as follows. Section 2 describes i-vector/PLDA framework as our baseline ASV system. The proposed method for i-vector extraction and modification for LDA are detailed in Sect. 3. Section 4 presents the experimental setups. Section 5 discusses system implementation and evaluation. Section 6 concludes the paper and outlines future studies.

2 Baseline ASV System Description

2.1 I-Vector Extraction

As mentioned earlier, i-vector based system has become de facto choice for speaker verification and related tasks. I-vector is essentially a low-dimensional representation of the Gaussian mixture model (GMM) super-vector found through a factor analysis process. Specifically, the speaker and channel dependent GMM super-vector M can be generated by

$$\mathrm{M} = \mathrm{m} + \mathrm{Tw} \tag{1}$$

where m is the speaker and channel independent super-vector, which is concatenated means of universal background model (UBM), T is a low-rank total variability (TV) matrix, and w is a random latent variable with standard normal distribution. In i-vector approach, the universal background model (UBM) and total variability (TV) matrix are trained with large amount speech data gathered from different speakers. The i-vector x is given by the maximum a posteriori (MAP) point estimate of the hidden variable w which is equal to the mean of the posterior distribution of w conditioned on input utterance:

$$\mathrm{x} = \left(I + T^T\Sigma^{-1}NT\right)^{-1}T^T\Sigma^{-1}N(E-m) \tag{2}$$

where Σ is a diagonal matrix, in which the diagonal blocks are corresponding covariance matrices of Gaussian components of the UBM, N and E are zero and first order Baum-Welch (BW) statistics matrices, respectively. Given an utterance $\mathrm{X} = \{x_1, x_2, \ldots, x_F\}$, the zero and first order BW statistics are computed using UBM as

$$N_i = \sum\nolimits_{j=1}^{F} Pr\left(i|x_j\right) \tag{3}$$

$$E_i(X) = \frac{1}{N_i}\sum\nolimits_{j=1}^{F} Pr\left(i|x_j\right)x_j \tag{4}$$

where $Pr\left(i|x_j\right)$ is posterior probability of generating x_j by corresponding Gaussian component density:

$$Pr\left(i|x_j\right) = \frac{\omega_i p_i\left(x_j\right)}{\Sigma_{k=1}^{C}\,\omega_k p_k\left(x_j\right)} \tag{5}$$

2.2 Linear Discriminant Analysis (LDA)

After the i-vector extraction, linear discriminant analysis (LDA) is used to compensate within-class variations and reduce the dimensionality prior to probabilistic linear discriminant analysis (PLDA) modeling. In LDA method, we simultaneously maximize the between-class variability and minimize the within-class variability by maximizing the following objective function:

$$\mathrm{J}(v) = \frac{v^T \Sigma_b v}{v^T \Sigma_w v} \tag{6}$$

where v is eigenvector, Σ_b and Σ_w are between-class scatter matrix and within-class scatter matrix, respectively, which are determined by

$$\Sigma_b = \sum\nolimits_{s=1}^{s} n_s(\bar{\mathrm{x}}_s - \bar{\mathrm{x}})(\bar{x}_s - \bar{x})^T \tag{7}$$

$$\Sigma_w = \sum\nolimits_{s=1}^{s} \sum\nolimits_{i=1}^{n_s} \left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)\left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)^T \tag{8}$$

where S is the number of all speakers, n_s is the number of utterances from speaker s, $\bar{\mathrm{X}}_s$ is the average of the i-vectors from speaker s, and $\bar{x}$ is the average of all i-vectors, defined as follows

$$\bar{\mathrm{x}}_s = \frac{1}{n_s} \sum\nolimits_{i=1}^{n_s} x_i^s \tag{9}$$

$$\bar{x} = \frac{1}{N} \sum\nolimits_{s=1}^{s} \sum\nolimits_{i=1}^{n_s} x_i^s \tag{10}$$

where N is the total number of utterances.

The LDA projection matrix is found by solving the following eigenvalue problem:

$$\Sigma_b v = \Lambda \Sigma_w v \tag{11}$$

where Λ is eigenvalue matrix. The projection matrix A is formalized by selecting first k eigenvectors corresponding to the k largest eigenvalues:

$$\mathrm{A} = \left[v_1, v_2 \ldots v_k\right] \tag{12}$$

Finally, the LDA compensated i-vectors are calculated as

$$x_{LDA} = A^T x \tag{13}$$

2.3 Probabilistic Linear Discriminant Analysis (PLDA)

Apart from compensating the within-class variations in i-vector space by subspace transformation, probabilistic linear discriminant analysis (PLDA) is widely used to reduce the redundant information such as channels from i-vectors. Here, the generative model for length-normalized i-vectors of s speaker with n_s sessions can be expressed as

$$\mathrm{x}_{\mathrm{i,j}} = \mu + \mathrm{V} z_i + \varepsilon_{i,j} \tag{14}$$

where μ is the mean of i-vectors, V defines the eigen-voice subspace, z_i is the speaker factor, and $\varepsilon_{i,j}$ is the residual term.

The verification scores of PLDA system is given as batch likelihood ratio. For projected enrollment and test i-vectors, z_{target} and z_{test}, the batch likelihood ratio is computed as

$$\Lambda(z_{target}, z_{test}) = \log \frac{p(z_{target}, z_{test}|H_1)}{p(z_{target}|H_0)p(z_{test}|H_1)} \tag{15}$$

where H_1 denotes the hypothesis that i-vectors belong to the same speaker and H_0 denotes the hypothesis that they are from different speakers. Figure 1 shows the process of calculating scores from the enrollment and test utterance in our i-vector/PLDA ASV system.

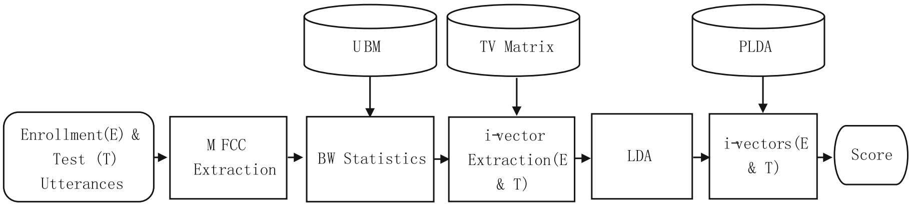


Fig. 1. Block diagram of i-vector/PLDA ASV system

3 Proposed System Modification

3.1 Analysis of I-Vector Estimation for Short Utterance

In i-vector systems, the test utterance and enrolment utterance(s) are represented by test and enrolment i-vectors extracted with pre-trained UBM and TV matrix. Then ASV is addressed by comparing the test i-vector with enrolment i-vector(s) signed by the individual to generate an accepted or rejected decision. Though the requirement of speech duration can somehow be met in enrolment stage, it may not be possible to maintain the same during the verification stage. This seriously limits the implementation of ASV system in real-world applications.

To better understand the effects of test duration variability on system performance, we present a detailed analysis of i-vector extraction pipeline. With short utterance, there is an increased uncertainty of BW statistics estimation due to lack of enough data to compute statistics parameters, which leads to an uncertain i-vector estimation. For i-vector systems, BW statistics totally represent the feature extracted from a test segment. [7, 11] Particularly, the zero-order BW statistics defines the covariance matrix of the posterior distribution given the utterance as

$$w_{\Sigma} = \left(I + T^{T}\Sigma^{-1}NT\right)^{-1} \tag{16}$$

where w_{Σ} is the covariance of the estimated i-vector, T is TV matrix, Σ is the UBM covariance, N is a diagonal matrix, where the diagonal blocks are the zero-order BW statistics of corresponding Gaussian components in UBM. Since the UBM and TV matrix are pre-trained with large quantity of data from different speakers, the higher variability introduced in BW statistics account for the uncertainty in i-vector estimation for short test segment.

3.2 Incorporating Historical Test Information in I-Vector Extraction

In order to improve the i-vector estimation, we propose a new method for adding historical test information in BW statistics computation. Rather than only use current test utterance to compute the BW statistics, we also exploit the weighted historical test utterance statistics to provide additional information. We define the weight γ_i as the estimated probability of current test utterance and historical test utterance i belonging to the same speaker. Then the BW statistics used to extract the current test i-vector is given by

$$N = N_c + \Sigma\gamma_i N_i \tag{17}$$

$$E = E_c + \Sigma\gamma_i E_i \tag{18}$$

where N_c and E_c are BW statistics computed from current test utterance, N_i and E_i are BW statistics computed from historical test utterance, and γ_i is corresponding weight assigned to historical test.

To compute the weight γ_i for historical test utterance, we use a world MAP estimator which was proposed in [12] and successfully applied to unsupervised GMM adaptation thereafter in [13, 14]. We first train a two-class Bayesian classifier based on two score models - target and non-target scores - learned from a development set. [14] Each score distribution is modelled by a 12 components GMM. Given the priori target and non-target score distributions, we can compute the posteriori probability of having a target. Specifically, for every encountered test utterance, ASV system output a raw score. Given current test raw score, s_0, the posteriori probability of this test belonging to the target speaker is defined as

$$P(tar|s_0) = \frac{P(s_0|tar)P_{tar}}{P(s_0|tar)P_{tar} + P(s_0|non)P_{non}} \tag{19}$$

where $P(s_0|tar)$ and $P(s_0|non)$ are the probabilities of the score given the target and non-target score distributions, P_{tar} and P_{non} are the prior probabilities of target and non-target test respectively. Then for historical test utterance i with raw score, s_i, we can compute weight γ_i as follows:

$$\gamma_i = P(tar|s_o)P(tar|s_i) + [1 - P(tar|s_0)][1 - P(tar|s_i)] \tag{20}$$

Note that all scores used are normalized. In proposed method, we do not require access to the historical test utterances as well as i-vectors. To utilize historical test information, only raw score and corresponding BW statistics are needed, which do not put a heavy burden on real-life applications. Figure 2 shows the flow diagram of the proposed method.

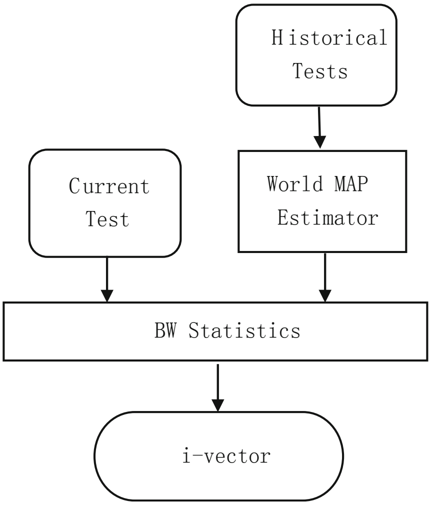


Fig. 2. Flow diagram of the proposed i-vector extraction method

3.3 Modified LDA for Domain Mismatch Compensation

One of the keys to the success of i-vector/PLDA framework is the use of a large quantity of previously collected speech data to characterize and model speaker and channel variability. However, it is unrealistic to assume such a large set of development data for every domain of interest. This is especially true for PLDA modeling, which needs labeled speech data, whereas the training of UBM and TV matrix only need unlabeled data. Studies have found that when PLDA is trained using out-domain data, the ASV system performance degrades rapidly due to the mismatch between development and evaluation data [15].

Conventional LDA projection falls to compensate this domain variability because it captures the domain variability in between-class scatter matrix. Instead of minimizing the domain mismatch in projected i-vectors, LDA maximizes domain variability when training the projection matrix. In order to address such problem, we modify the LDA training to separate domain variability from scatter matrix estimation. For simplicity, we assume the speakers do not overlap across different domains. In our method, the new between-class scatter matrix and within-class scatter matrix are defined as

$$\Sigma'_b = \sum\nolimits_{s=1}^{S_{OUT}} n_s(\bar{\mathrm{x}}_s - \bar{\mathrm{x}}_{out})(\bar{\mathrm{x}}_s - \bar{\mathrm{x}}_{out})^T + \sum\nolimits_{s=1}^{s_{in}} n_s(\bar{\mathrm{x}}_s - \bar{\mathrm{x}}_{out})(\bar{\mathrm{x}}_s - \bar{\mathrm{x}}_{out})^T \tag{21}$$

$$\Sigma'_w = \sum\nolimits_{s=1}^{S_{OUT}} \sum\nolimits_{i=1}^{n_s} \left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)\left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)^T + \sum\nolimits_{s=1}^{s_{in}} \sum\nolimits_{i=1}^{n_s} \left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)\left(\mathrm{x}_i^s - \bar{\mathrm{x}}_s\right)^T \tag{22}$$

where S_{out} and S_{in} are the number of out-domain and in-domain speakers, $\bar{\mathrm{x}}_{out}$ and $\bar{\mathrm{x}}_{in}$ are the average of the out-domain and in-domain i-vectors, respectively. Also, we define inter-domain variability matrix as

$$\Sigma_d = S_{out}(\bar{\mathrm{x}}_{out} - \bar{x})(\bar{\mathrm{x}}_{out} - \bar{x})^T + S_{in}(\bar{\mathrm{x}}_{in} - \bar{x})(\bar{\mathrm{x}}_{in} - \bar{x})^T \tag{23}$$

Finally, the modified LDA projection matrix can be calculated by maximizing the following objective function,

$$\mathrm{J}(v) = \frac{v^T \sum\nolimits'_b v}{v^T \sum\nolimits_{wd} v} \tag{24}$$

where v is eigenvector, and $\Sigma_{wd} = \Sigma'_w \Sigma_d^T$. By maximizing above objective function, we can simultaneously maximize the between-class variability and minimizing both within-class variability and domain variability.

4 Experimental Setups

4.1 Speech Data and Acoustic Features

Audio data are collected by an integrated microphone from power grid dispatching hall and dispatcher training simulator (DTS) room. All speakers are male. The raw data are automatically saved in a memory card every 3 min. The two locations have different room sizes, background noises, telephone channels, and so on. Figure 3 shows different environmental setting of audio data collection. From raw audio data, 19 dimensional Mel-frequency cepstral coefficients (MFCCs) together with energy coefficient are extracted and appended with delta and delta-delta features to form a 60-dimensional vector. The vector is extracted every 10 ms, using a Hamming window of 20 ms. And silence frames are detected and discarded by an energy-based voice activity detector (VAD).

(a)

(b)

Fig. 3. Different locations of audio data collection. (a) Power grid dispatching hall. (b) DTS room

Unless stated otherwise, we partition data gathered from DTS room into two subsets. We use one subset as development data and the other as evaluation data. In order to carry out experiments for short utterance conditions, original speech utterances are split into 2 s, 5 s, 10 s (only contain active frames) duration as short test segments. We randomly select initial frame and create 500 truncated segments for each duration. To test the effectiveness of modified LDA in ASV tasks with limited target domain data, we frame the domain mismatch compensation problem as reducing the mismatch between the data collected from different locations. We regard speech utterances collected from power grid dispatching hall as in-domain data, and utterances collected from dispatcher training simulator (DTS) room are considered as out-domain data. In this case, the speech files from DTS room are used as development data and speech files from power grid dispatching hall are used as evaluation data.

4.2 I-Vector Extraction and PLDA Modeling

To extract i-vector, we train a UBM with 512 Gaussian components on development data and use UBM to estimate the BW statistics. The TV subspace has a dimension of 400 and is trained on same development data. For LDA and modified LDA training, the reduced dimension is kept at 200. Length normalization is applied to LDA projected i-vectors to convert their behavior into Gaussian. Then a PLDA model with 150 latent variables is trained. We train the World MAP estimator on development data. The prior probability used are 0.1 for target and 0.9 for non-target.

4.3 Evaluation Criteria

There are two kind of mistakes in ASV system: a false rejection happens when a genuine speaker is incorrectly rejected and a false alarm when an imposter is accepted. In our experiment, the system performance is evaluated using equal error rate (EER) in which the false rejection rate and false alarm rate are equal. Also, we report experimental results in terms of minimum detection cost function (minDCF).

5 Results and Discussions

5.1 Baseline ASV System Performance

In the first series of experiments, we compare the performance of baseline ASV system in different test durations. The experiments are conducted on speech files collected from DTS room. We use 3 min raw speech for enrollment and three types of truncated segments (contain 2 s, 5 s 10 s active frames respectively) for test i-vector extraction. The results are presented in Fig. 4.

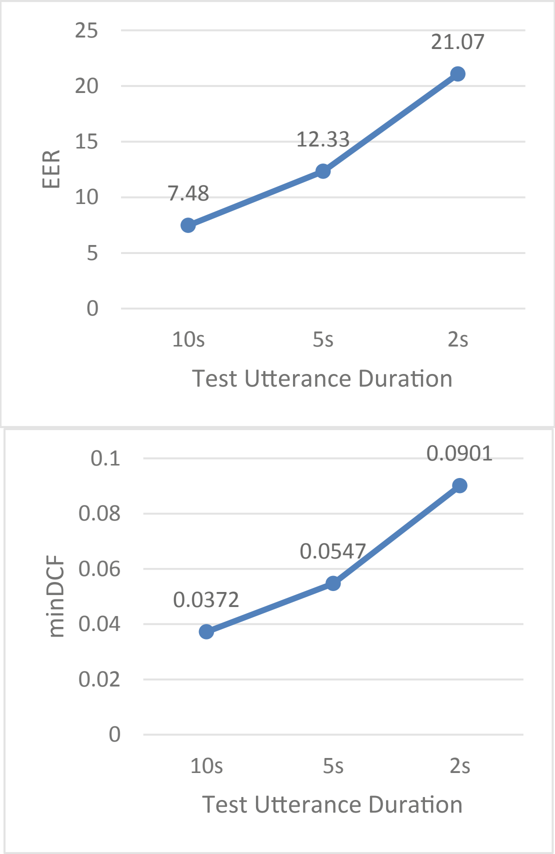


Fig. 4. Baseline ASV system performance for different test duration conditions

It can be observed that system performance in terms of both EER and minDCF degrades monotonically with the decrease in speech duration. When ASV system is presented with 2 s short utterance, the EER and minDCF increase 182% and 142% respectively compared to 10 s test utterance. This illustrates the need for proposed i-vector extraction method.

Next, we use speech files from power grid dispatching hall as evaluation data. Similarly, 3 min raw speech is used for enrollment and 2 s, 5 s, 10 s truncated speech segments are used for testing. This series of experiments aims to show the effect of in-domain development data on the performance of baseline system. The results are presented in Fig. 5.

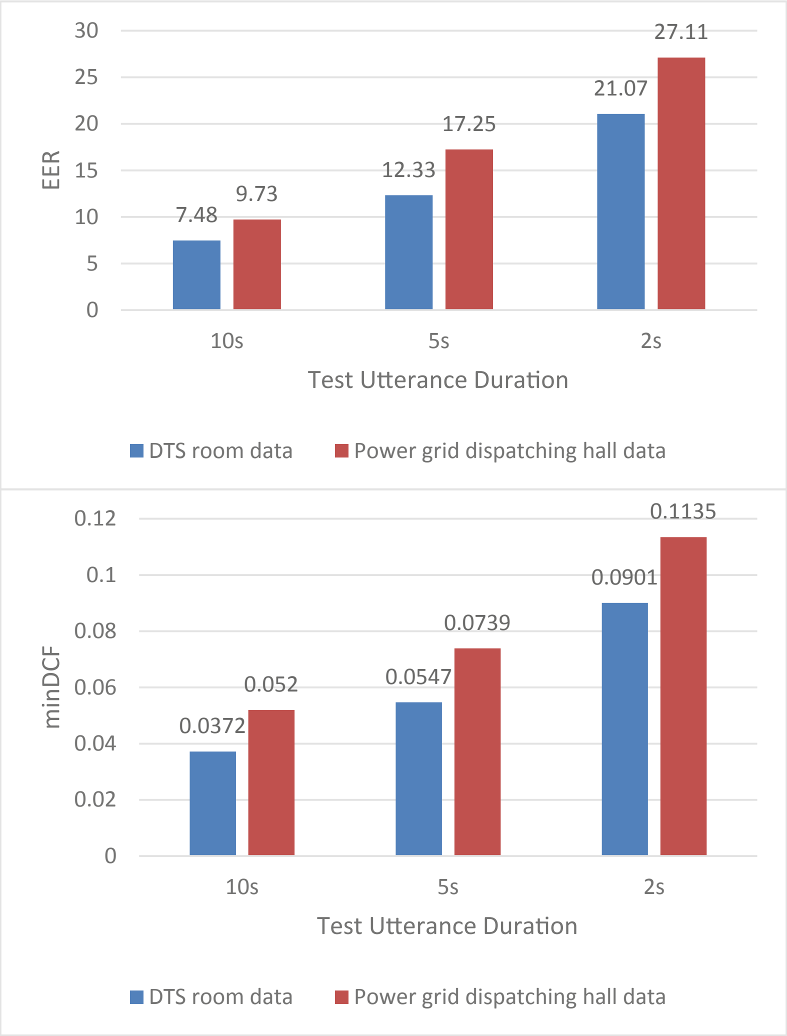


Fig. 5. Performance comparison using in-domain and out-domain development data

As shown in Fig. 5, there is a gap in performance on power grid dispatching hall enroll/test set when hyper-parameters are trained with development data gathered from DTS room. In Sect. 5.3, we employ modified LDA to reduce this performance degradation.

5.2 Proposed Method for I-Vector Extraction

In this section, we conduct experiments to test the effectiveness of incorporating historical test information in short utterance i-vector extraction. We use speech files collected from DTS room as both development and evaluation data. The results are presented in Table 1.

Table 1. Performance comparison of baseline system and system incorporating historical test information (proposed-1) in i-vector extraction

	EER			minDCF		
	10 s	5 s	2 s	10 s	5 s	2 s
Baseline (matched)	7.48	12.33	21.07	0.0372	0.0547	0.0901
Proposed-1	7.09	11.45	19.08	0.0359	0.0518	0.0827
Relative improvement	5.2%	7.1%	9.4%	3.4%	5.3%	8.2%

Experimental results reported in Table 1 show when enough historical information is inserted, the proposed method could achieve noticeable improvement in terms of EER and minDCF compared with the baseline i-vector system in different short duration conditions. We observe that the relative improvement increases with the decrease in test utterance duration. This suggests that incorporating historical information is useful for short utterance.

To analyze the behavior of our method more precisely, we investigate the system performance in terms of EER for each newly added test utterance. We conduct the experiment on 10 random draws from the entire truncated speech segments pool and

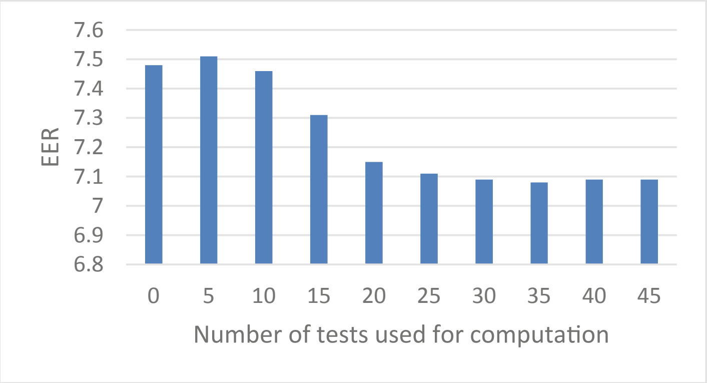


Fig. 6. Average EER of the 10 s test utterance condition

evaluate the performance individually. The results are averaged over 10 random draws for statistical significance. We notice that a minimum amount of data should be presented for proposed system to obtain stable gain. The average EER of 10 s test utterance condition are presented in Fig. 6. In 2 s and 5 s utterance conditions, the patterns are similar.

5.3 Modified LDA

As shown in Sect. 5.1, when ASV system is developed using data which is outside the target domain, it significantly affects the performance due to the mismatch between development and evaluation data. To investigate this situation, we use speech files collected from DTS room as development data and speech files collected from power grid dispatching hall as evaluation data. We use modified LDA projection to replace the conventional LDA in baseline system. System performance in terms of EER and minDCF are presented in the Table 2.

Table 2. Performance comparison of baseline system, system with modified LDA (proposed-2)

	EER			minDCF		
	10 s	5 s	2 s	10 s	5 s	2 s
Baseline (matched)	7.48	12.33	21.07	0.0372	0.0547	0.0901
Baseline (mismatched)	9.73	17.25	27.11	0.052	0.0739	0.1135
Proposed-2	7.74	14.15	22.67	0.0398	0.0588	0.0927

From Table 2, a relative gain of at least 16.4% in EER and 18.3% in minDCF is observed after applying modified LDA. In terms of bridging the performance gap between a matched baseline (DTS room data for both development and evaluation) and a mismatched baseline (DTS room data for development, power grid dispatching hall data for evaluation) system, we are able to recover at least 63% of the performance gap for different duration conditions. It demonstrates that modified LDA is quite successful in reducing the volume of in-domain development data.

Finally, we conduct experiment on system integrating the proposed i-vector extraction method and modified LDA. We develop system on speech data collected from DTS room and evaluate performance on data collected from power grid dispatching hall. From Table 3, it can be observed that further improvement is achieved with combined approach. Compared to baseline, it shows at least 20% improvement for different test segment durations.

Table 3. Performance of system using combined approach (proposed-3)

	EER			minDCF		
	10 s	5 s	2 s	10 s	5 s	2 s
Baseline (mismatched)	9.73	17.25	27.11	0.052	0.0739	0.1135
Proposed-3	7.3	13.19	21.63	0.0367	0.0546	0.0903

6 Conclusions and Future Work

The performance of i-vector/PLDA ASV systems depends on a large quantity of in-domain development data for PLDA training. During the evaluation, it is also critical that the speech duration is long enough to reduce the uncertainty in i-vector estimation. In many practical applications, the speaker verification performance is affected due to the difficulty in collecting significant amount of speech data. In this study, we propose modification for i-vector ASV system to address the issue of performance degradation with limited voice data. With the aid of historical test information, we observe a relative improvement of 9.4% in EER for 2 s test duration condition. When system is trained on mismatched development dataset, we are able to recover at least 63% of performance gap using modified LDA projection. The best performance is achieved with combined method, where we obtain relative improvement in the range of 20–29% over baseline system.

Despite the promising results, there are still some problems to study in the future. For example, currently world MAP estimator assumes the prior probabilities when the corresponding scores are not encountered in the score GMM training data. While it is anticipated that this situation is rare, we intend to investigate its effect on system performance. In addition, speakers can overlap in different domains and the data in one domain can be multi-modal. Such multi-modality can lead to misrepresentation of the speaker and non-speaker information [16]. We intend to extend our modified LDA method to compensate for speaker population difference among different portions of training data. Also, we intend to investigate the relationship between system performance and different sizes of in-domain data used for LDA training. In our future work, we intend to explore applying proposed methods onto deep neural networks (DNN) based systems. Using DNN instead of GMM to derive speaker specific information is a very promising direction to look at.

References

1. Yang, F., Wu, C., Wang, F., et al.: Review of studies on human reliability researches during 1998 to 2008. Sci. Technol. Rev. **27**(8), 87–94 (2009)
2. Dehak, N., Kenny, P., Dehak, R., et al.: Front-end factor analysis for speaker verification. IEEE/ACM Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
3. Cai, W., Li, M., Li, L., et al.: Duration dependent covariance regularization in PLDA modeling for speaker verification. In: Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1027–1031 (2015)
4. Kanagasundaram, A., Dean, D., Sridharan, S., et al.: A study on the effects of using short utterance length development data in the design of GPLDA speaker verification systems. Int. J. Speech Technol. **20**(2), 247–259 (2017)
5. Kenny, P., Stafylakis, T., Ouellet, P., et al.: PLDA for speaker verification with utterances of arbitrary duration. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7649–7653 (2013)
6. Cumani, S., Plchot, O., Laface, P.: On the use of i-vector posterior distributions in probabilistic linear discriminant analysis. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(4), 846–857 (2014)

7. Poddar, A., Sahidullah, M., Saha, G.: Improved i-vector extraction technique for speaker verification with short utterances. Int. J. Speech Technol. **21**(3), 473–488 (2018)
8. Villalba, J., Lleida, E.: Unsupervised adaptation of PLDA by using variational bayes methods. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 744–748 (2014)
9. Shum, S., Reynolds, D.A., Garcia-Romero, D., et al.: Unsupervised clustering approaches for domain adaptation in speaker recognition systems. In: Odyssey: The Speaker and Language Recognition Workshop, pp. 266–272 (2014)
10. Aronowitz, H.: Compensating inter-dataset variability in PLDA hyper-parameters for robust speaker recognition. In: Odyssey: The Speaker and Language Recognition Workshop, pp. 280–286 (2014)
11. Kenny, P., Ouellet, P., Dehak, N., et al.: A study of interspeaker variability in speaker verification. IEEE/ACM Trans. Audio Speech Lang. Process. **16**(5), 980–988 (2008)
12. Fredouille, C., Bonastre, J.F., Merlin, T.: Bayesian approach-based decision in speaker verification. In: Odyssey: The Speaker and Language Recognition Workshop, pp. 77–81 (2001)
13. Preti, A., Bonastre, J.F., Matrouf, D.: Confidence measure based unsupervised target model adaptation for speaker verification. In: Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 754–757 (2007)
14. Mclaren, M., Matrouf, D., Vogt, R., et al.: Applying SVMs and weight-based factor analysis to unsupervised adaptation for speaker verification. Comput. Speech Lang. **25**(2), 327–340 (2011)
15. Rahman, M.H., Kanagasundaram, A., Himawan, I., et al.: Improving PLDA speaker verification performance using domain mismatch compensation techniques. Comput. Speech Lang. **47**, 240–258 (2017)
16. Glembek, O., Ma, J., Matejka, P., et al.: Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4060–4064 (2014)

Correction to: Impacts of Automation Reliability and Failure Modes on Operators' Performance in Security Screening

Zijian Yin, Pei-Luen Patrick Rau, and Zhizhong Li

Correction to:
Chapter "Impacts of Automation Reliability and Failure Modes on Operators' Performance in Security Screening" in: D. Harris (Ed.): *Engineering Psychology and Cognitive Ergonomics*, LNAI 11571, https://doi.org/10.1007/978-3-030-22507-0_11

In the original version of this chapter Reference 6 was published incorrectly. Reference 6 has now been corrected.

The updated version of this chapter can be found at
https://doi.org/10.1007/978-3-030-22507-0_11

D. Harris (Ed.): HCII 2019, LNAI 11571, p. C1, 2020.
https://doi.org/10.1007/978-3-030-22507-0_34

Author Index

Barber, Daniel 363
Behrendt, Marcus 187
Blackie, Andrew 76
Boy, Guy A. 321
Braithwaite, Graham 66

Chassin, David P. 39

Degen, Helmut 3
Deng, Ye 20

E, Xiaotian 150

Franssila, Heljä 29
Fu, Shan 92, 167, 271, 423

Gao, Xiaotong 396
Guo, Junpeng 249

Hattori, Yuki 410
Hidalgo, Maartje 363
Hildebrandt, Michael 55
Hou, Wenjun 176
Hou, Yiping 200
Hu, Wan-Lin 39
Huang, Wenyan 55
Huangfu, Huayan 167, 271

Jakubowski, Julius 259
Jiang, Zhiyang 176
Jie, Yuwen 353
Josefsson, Billy 381

Kearney, Peter 66
Knoop, Dennis 187
Komandur, Sashidharan 106
Kong, Xiangjie 249

Lau, Nathan 55
Li, Jiaxuan 236
Li, Ming 200
Li, Wen-Chin 66, 259
Li, Zhizhong 20, 137
Liu, Yuqing 249
Liu, Zhongqi 236

Lu, Yanyu 92, 423
Lu, Yi 167, 271
Luengo, Sven Malmberg 381
Lundberg, Jonas 381

MacDonald, Erin 39
McCarthy, Pete 76
Meyer, Lothar 381
Miyake, Yoshihiro 410

Nozawa, Takayuki 410

Peng, Caifang 92
Peukert, Maximilian 381
Pignoni, Giovanni 106

Qian, Ping 308
Qiao, Han 308
Qin, Hua 227, 396
Qiu, Kai 214

Ran, Linghua 227
Rau, Pei-Luen Patrick 137
Reinerman-Jones, Lauren 363
Rivetta, Claudio 39
Roijendijk, Linsey 123

Sarinasadat, Hosseini 410
Sevillian, Dujuan B. 284
She, Manrong 20
Song, Yang 308
Su, Jianning 214
Sun, Xianghong 308
Sun, Yanjin 308
Svensson, Åsa 381

Tan, Wei 321

van Beurden, Maurice 123

Wang, Lei 342
Wang, Yuexuan 20
Wang, Zhen 92, 423
Wang, Zheng 423

Wortelen, Bertram 187
Wu, Lei 332

Xu, Jian 332
Xu, Yunshu 55

Yang, Jin 249
Yang, Wenjin 214
Yin, Zijian 137
Yu, Saisai 150
Yue, Chen 227

Zhang, Jie 200
Zhang, Jingyi 342
Zhang, Jingyu 150, 308
Zhang, Shuguang 271
Zhang, Shutao 214
Zhang, Xinxin 214
Zhang, Yijing 20
Zheng, Yiyuan 353
Zhou, Qianxiang 236
Zhou, Shihua 249