Don Harris (Ed.)

LNAI 8532

Engineering Psychology and Cognitive Ergonomics

11th International Conference, EPCE 2014
Held as Part of HCI International 2014
Heraklion, Crete, Greece, June 22–27, 2014, Proceedings

Springer

Lecture Notes in Artificial Intelligence 8532

Subseries of Lecture Notes in Computer Science

Don Harris (Ed.)

Engineering Psychology and Cognitive Ergonomics

11th International Conference, EPCE 2014
Held as Part of HCI International 2014
Heraklion, Crete, Greece, June 22-27, 2014
Proceedings

Springer

Volume Editor

Don Harris
Coventry University
Faculty of Engineering and Computing
Priory Street
Coventry CV1 5FB, UK
E-mail: don.harris@coventry.ac.uk

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-07514-3 e-ISBN 978-3-319-07515-0
DOI 10.1007/978-3-319-07515-0
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014939567

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The 16th International Conference on Human–Computer Interaction, HCI International 2014, was held in Heraklion, Crete, Greece, during June 22–27, 2014, incorporating 14 conferences/thematic areas:

Thematic areas:

- Human–Computer Interaction
- Human Interface and the Management of Information

Affiliated conferences:

- 11th International Conference on Engineering Psychology and Cognitive Ergonomics
- 8th International Conference on Universal Access in Human–Computer Interaction
- 6th International Conference on Virtual, Augmented and Mixed Reality
- 6th International Conference on Cross-Cultural Design
- 6th International Conference on Social Computing and Social Media
- 8th International Conference on Augmented Cognition
- 5th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management
- Third International Conference on Design, User Experience and Usability
- Second International Conference on Distributed, Ambient and Pervasive Interactions
- Second International Conference on Human Aspects of Information Security, Privacy and Trust
- First International Conference on HCI in Business
- First International Conference on Learning and Collaboration Technologies

A total of 4,766 individuals from academia, research institutes, industry, and governmental agencies from 78 countries submitted contributions, and 1,476 papers and 225 posters were included in the proceedings. These papers address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The papers 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.

This volume, edited by Don Harris, contains papers focusing on the thematic area of engineering psychology and cognitive ergonomics, addressing the following major topics:

- Mental workload and stress
- Visual Perception

- Cognitive issues in interaction and user experience
- Cognitive Psychology in aviation and space
- Transport and industrial applications

The remaining volumes of the HCI International 2014 proceedings are:

- Volume 1, LNCS 8510, Human–Computer Interaction: HCI Theories, Methods and Tools (Part I), edited by Masaaki Kurosu
- Volume 2, LNCS 8511, Human–Computer Interaction: Advanced Interaction Modalities and Techniques (Part II), edited by Masaaki Kurosu
- Volume 3, LNCS 8512, Human–Computer Interaction: Applications and Services (Part III), edited by Masaaki Kurosu
- Volume 4, LNCS 8513, Universal Access in Human–Computer Interaction: Design and Development Methods for Universal Access (Part I), edited by Constantine Stephanidis and Margherita Antona
- Volume 5, LNCS 8514, Universal Access in Human–Computer Interaction: Universal Access to Information and Knowledge (Part II), edited by Constantine Stephanidisand MargheritaAntona
- Volume 6, LNCS 8515, Universal Access in Human–Computer Interaction: Aging and Assistive Environments (Part III), edited by Constantine Stephanidis and Margherita Antona
- Volume 7, LNCS 8516, Universal Access in Human–Computer Interaction: Design for All and Accessibility Practice (Part IV), edited by Constantine Stephanidis and Margherita Antona
- Volume 8, LNCS 8517, Design, User Experience, and Usability: Theories, Methods and Tools for Designing the User Experience (Part I), edited by Aaron Marcus
- Volume 9, LNCS 8518, Design, User Experience, and Usability: User Experience Design for Diverse Interaction Platforms and Environments (Part II), edited by Aaron Marcus
- Volume 10, LNCS 8519, Design, User Experience, and Usability: User Experience Design for Everyday Life Applications and Services (Part III), edited by Aaron Marcus
- Volume 11, LNCS 8520, Design, User Experience, and Usability: User Experience Design Practice (Part IV), edited by Aaron Marcus
- Volume 12, LNCS 8521, Human Interface and the Management of Information: Information and Knowledge Design and Evaluation (Part I), edited by Sakae Yamamoto
- Volume 13, LNCS 8522, Human Interface and the Management of Information: Information and Knowledge in Applications and Services (Part II), edited by Sakae Yamamoto
- Volume 14, LNCS 8523, Learning and Collaboration Technologies: Designing and Developing Novel Learning Experiences (Part I), edited by Panayiotis Zaphiris and Andri Ioannou
- Volume 15, LNCS 8524, Learning and Collaboration Technologies: Technology-rich Environments for Learning and Collaboration (Part II), edited by Panayiotis Zaphiris and Andri Ioannou

- Volume 16, LNCS 8525, Virtual, Augmented and Mixed Reality: Designing and Developing Virtual and Augmented Environments (Part I), edited by Randall Shumaker and Stephanie Lackey
- Volume 17, LNCS 8526, Virtual, Augmented and Mixed Reality: Applications of Virtual and Augmented Reality (Part II), edited by Randall Shumaker and Stephanie Lackey
- Volume 18, LNCS 8527, HCI in Business, edited by Fiona Fui-Hoon Nah
- Volume 19, LNCS 8528, Cross-Cultural Design, edited by P.L. Patrick Rau
- Volume 20, LNCS 8529, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management, edited by Vincent G. Duffy
- Volume 21, LNCS 8530, Distributed, Ambient, and Pervasive Interactions, edited by Norbert Streitz and Panos Markopoulos
- Volume 22, LNCS 8531, Social Computing and Social Media, edited by Gabriele Meiselwitz
- Volume 24, LNCS 8533, Human Aspects of Information Security, Privacy and Trust, edited by Theo Tryfonasand IoannisAskoxylakis
- Volume 25, LNAI 8534, Foundations of Augmented Cognition, edited by Dylan D. Schmorrow and Cali M. Fidopiastis
- Volume 26, CCIS 434, HCI International 2014 Posters Proceedings (Part I), edited by Constantine Stephanidis
- Volume 27, CCIS 435, HCI International 2014 Posters Proceedings (Part II), edited by Constantine Stephanidis

I would like to thank the Program Chairs and the members of the Program Boards of all affiliated conferences and thematic areas, listed below, for their contribution to the highest scientific quality and the overall success of the HCI International 2014 Conference.

This conference could not have been possible without the continuous support and advice of the founding chair and conference scientific advisor, Prof. Gavriel Salvendy, as well as the dedicated work and outstanding efforts of the communications chair and editor of *HCI International News*, Dr. Abbas Moallem.

I would also like to thank for their contribution towards the smooth organization of the HCI International 2014 Conference the members of the Human–Computer Interaction Laboratory of ICS-FORTH, and in particular George Paparoulis, Maria Pitsoulaki, Maria Bouhli, and George Kapnas.

April 2014

Constantine Stephanidis
General Chair, HCI International 2014

Organization

Human–Computer Interaction

Program Chair: Masaaki Kurosu, Japan

Jose Abdelnour-Nocera, UK
Sebastiano Bagnara, Italy
Simone Barbosa, Brazil
Adriana Betiol, Brazil
Simone Borsci, UK
Henry Duh, Australia
Xiaowen Fang, USA
Vicki Hanson, UK
Wonil Hwang, Korea
Minna Isomursu, Finland
Yong Gu Ji, Korea
Anirudha Joshi, India
Esther Jun, USA
Kyungdoh Kim, Korea
Heidi Krömker, Germany
Chen Ling, USA
Chang S. Nam, USA
Naoko Okuizumi, Japan
Philippe Palanque, France
Ling Rothrock, USA
Naoki Sakakibara, Japan
Dominique Scapin, France
Guangfeng Song, USA
Sanjay Tripathi, India
Chui Yin Wong, Malaysia
Toshiki Yamaoka, Japan
Kazuhiko Yamazaki, Japan
Ryoji Yoshitake, Japan

Human Interface and the Management of Information

Program Chair: Sakae Yamamoto, Japan

Alan Chan, Hong Kong
Denis A. Coelho, Portugal
Linda Elliott, USA
Shin'ichi Fukuzumi, Japan
Michitaka Hirose, Japan
Makoto Itoh, Japan
Yen-Yu Kang, Taiwan
Koji Kimita, Japan
Daiji Kobayashi, Japan
Hiroyuki Miki, Japan
Hirohiko Mori, Japan
Shogo Nishida, Japan
Robert Proctor, USA
Youngho Rhee, Korea
Ryosuke Saga, Japan
Katsunori Shimohara, Japan
Kim-Phuong Vu, USA
Tomio Watanabe, Japan

Engineering Psychology and Cognitive Ergonomics

Program Chair: Don Harris, UK

Guy Andre Boy, USA
Shan Fu, P.R. China
Hung-Sying Jing, Taiwan
Wen-Chin Li, Taiwan
Mark Neerincx, The Netherlands
Jan Noyes, UK
Paul Salmon, Australia
Axel Schulte, Germany
Siraj Shaikh, UK
Sarah Sharples, UK
Anthony Smoker, UK
Neville Stanton, UK
Alex Stedmon, UK
Andrew Thatcher, South Africa

Universal Access in Human–Computer Interaction

Program Chairs: Constantine Stephanidis, Greece, and Margherita Antona, Greece

Julio Abascal, Spain
Gisela Susanne Bahr, USA
João Barroso, Portugal
Margrit Betke, USA
Anthony Brooks, Denmark
Christian Bühler, Germany
Stefan Carmien, Spain
Hua Dong, P.R. China
Carlos Duarte, Portugal
Pier Luigi Emiliani, Italy
Qin Gao, P.R. China
Andrina Granić, Croatia
Andreas Holzinger, Austria
Josette Jones, USA
Simeon Keates, UK
Georgios Kouroupetroglou, Greece
Patrick Langdon, UK
Barbara Leporini, Italy
Eugene Loos, The Netherlands
Ana Isabel Paraguay, Brazil
Helen Petrie, UK
Michael Pieper, Germany
Enrico Pontelli, USA
Jaime Sanchez, Chile
Alberto Sanna, Italy
Anthony Savidis, Greece
Christian Stary, Austria
Hirotada Ueda, Japan
Gerhard Weber, Germany
Harald Weber, Germany

Virtual, Augmented and Mixed Reality

Program Chairs: Randall Shumaker, USA, and Stephanie Lackey, USA

Roland Blach, Germany
Sheryl Brahnam, USA
Juan Cendan, USA
Jessie Chen, USA
Panagiotis D. Kaklis, UK
Hirokazu Kato, Japan
Denis Laurendeau, Canada
Fotis Liarokapis, UK
Michael Macedonia, USA
Gordon Mair, UK

Jose San Martin, Spain
Tabitha Peck, USA
Christian Sandor, Australia
Christopher Stapleton, USA
Gregory Welch, USA

Cross-Cultural Design

Program Chair: P.L. Patrick Rau, P.R. China

Yee-Yin Choong, USA
Paul Fu, USA
Zhiyong Fu, P.R. China
Pin-Chao Liao, P.R. China
Dyi-Yih Michael Lin, Taiwan
Rungtai Lin, Taiwan
Ta-Ping (Robert) Lu, Taiwan
Liang Ma, P.R. China
Alexander Mädche, Germany
Sheau-Farn Max Liang, Taiwan
Katsuhiko Ogawa, Japan
Tom Plocher, USA
Huatong Sun, USA
Emil Tso, P.R. China
Hsiu-Ping Yueh, Taiwan
Liang (Leon) Zeng, USA
Jia Zhou, P.R. China

Online Communities and Social Media

Program Chair: Gabriele Meiselwitz, USA

Leonelo Almeida, Brazil
Chee Siang Ang, UK
Aneesha Bakharia, Australia
Ania Bobrowicz, UK
James Braman, USA
Farzin Deravi, UK
Carsten Kleiner, Germany
Niki Lambropoulos, Greece
Soo Ling Lim, UK
Anthony Norcio, USA
Portia Pusey, USA
Panote Siriaraya, UK
Stefan Stieglitz, Germany
Giovanni Vincenti, USA
Yuanqiong (Kathy) Wang, USA
June Wei, USA
Brian Wentz, USA

Augmented Cognition

Program Chairs: Dylan D. Schmorrow, USA, and Cali M. Fidopiastis, USA

Ahmed Abdelkhalek, USA
Robert Atkinson, USA
Monique Beaudoin, USA
John Blitch, USA
Alenka Brown, USA
Rosario Cannavò, Italy
Joseph Cohn, USA
Andrew J. Cowell, USA
Martha Crosby, USA
Wai-Tat Fu, USA

Rodolphe Gentili, USA
Frederick Gregory, USA
Michael W. Hail, USA
Monte Hancock, USA
Fei Hu, USA
Ion Juvina, USA
Joe Keebler, USA
Philip Mangos, USA
Rao Mannepalli, USA
David Martinez, USA
Yvonne R. Masakowski, USA
Santosh Mathan, USA
Ranjeev Mittu, USA
Keith Niall, USA
Tatana Olson, USA
Debra Patton, USA
June Pilcher, USA
Robinson Pino, USA
Tiffany Poeppelman, USA
Victoria Romero, USA
Amela Sadagic, USA
Anna Skinner, USA
Ann Speed, USA
Robert Sottilare, USA
Peter Walker, USA

Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management

Program Chair: Vincent G. Duffy, USA

Giuseppe Andreoni, Italy
Daniel Carruth, USA
Elsbeth De Korte, The Netherlands
Afzal A. Godil, USA
Ravindra Goonetilleke, Hong Kong
Noriaki Kuwahara, Japan
Kang Li, USA
Zhizhong Li, P.R. China
Tim Marler, USA
Jianwei Niu, P.R. China
Michelle Robertson, USA
Matthias Rötting, Germany
Mao-Jiun Wang, Taiwan
Xuguang Wang, France
James Yang, USA

Design, User Experience, and Usability

Program Chair: Aaron Marcus, USA

Sisira Adikari, Australia
Claire Ancient, USA
Arne Berger, Germany
Jamie Blustein, Canada
Ana Boa-Ventura, USA
Jan Brejcha, Czech Republic
Lorenzo Cantoni, Switzerland
Marc Fabri, UK
Luciane Maria Fadel, Brazil
Tricia Flanagan, Hong Kong
Jorge Frascara, Mexico
Federico Gobbo, Italy
Emilie Gould, USA
Rüdiger Heimgärtner, Germany
Brigitte Herrmann, Germany
Steffen Hess, Germany
Nouf Khashman, Canada
Fabiola Guillermina Noël, Mexico
Francisco Rebelo, Portugal
Kerem Rızvanoğlu, Turkey
Marcelo Soares, Brazil
Carla Spinillo, Brazil

Distributed, Ambient and Pervasive Interactions

Program Chairs: Norbert Streitz, Germany, and Panos Markopoulos, The Netherlands

Juan Carlos Augusto, UK
Jose Bravo, Spain
Adrian Cheok, UK
Boris de Ruyter, The Netherlands
Anind Dey, USA
Dimitris Grammenos, Greece
Nuno Guimaraes, Portugal
Achilles Kameas, Greece
Javed Vassilis Khan, The Netherlands
Shin'ichi Konomi, Japan
Carsten Magerkurth, Switzerland
Ingrid Mulder, The Netherlands
Anton Nijholt, The Netherlands
Fabio Paternó, Italy
Carsten Röcker, Germany
Teresa Romao, Portugal
Albert Ali Salah, Turkey
Manfred Tscheligi, Austria
Reiner Wichert, Germany
Woontack Woo, Korea
Xenophon Zabulis, Greece

Human Aspects of Information Security, Privacy and Trust

Program Chairs: Theo Tryfonas, UK, and Ioannis Askoxylakis, Greece

Claudio Agostino Ardagna, Italy
Zinaida Benenson, Germany
Daniele Catteddu, Italy
Raoul Chiesa, Italy
Bryan Cline, USA
Sadie Creese, UK
Jorge Cuellar, Germany
Marc Dacier, USA
Dieter Gollmann, Germany
Kirstie Hawkey, Canada
Jaap-Henk Hoepman, The Netherlands
Cagatay Karabat, Turkey
Angelos Keromytis, USA
Ayako Komatsu, Japan
Ronald Leenes, The Netherlands
Javier Lopez, Spain
Steve Marsh, Canada
Gregorio Martinez, Spain
Emilio Mordini, Italy
Yuko Murayama, Japan
Masakatsu Nishigaki, Japan
Aljosa Pasic, Spain
Milan Petković, The Netherlands
Joachim Posegga, Germany
Jean-Jacques Quisquater, Belgium
Damien Sauveron, France
George Spanoudakis, UK
Kerry-Lynn Thomson, South Africa
Julien Touzeau, France
Theo Tryfonas, UK
João Vilela, Portugal
Claire Vishik, UK
Melanie Volkamer, Germany

HCI in Business

Program Chair: Fiona Fui-Hoon Nah, USA

Andreas Auinger, Austria
Michel Avital, Denmark
Traci Carte, USA
Hock Chuan Chan, Singapore
Constantinos Coursaris, USA
Soussan Djamasbi, USA
Brenda Eschenbrenner, USA
Nobuyuki Fukawa, USA
Khaled Hassanein, Canada
Milena Head, Canada
Susanna (Shuk Ying) Ho, Australia
Jack Zhenhui Jiang, Singapore
Jinwoo Kim, Korea
Zoonky Lee, Korea
Honglei Li, UK
Nicholas Lockwood, USA
Eleanor T. Loiacono, USA
Mei Lu, USA
Scott McCoy, USA
Brian Mennecke, USA
Robin Poston, USA
Lingyun Qiu, P.R. China
Rene Riedl, Austria
Matti Rossi, Finland
April Savoy, USA
Shu Schiller, USA
Hong Sheng, USA
Choon Ling Sia, Hong Kong
Chee-Wee Tan, Denmark
Chuan Hoo Tan, Hong Kong
Noam Tractinsky, Israel
Horst Treiblmaier, Austria
Virpi Tuunainen, Finland
Dezhi Wu, USA
I-Chin Wu, Taiwan

Learning and Collaboration Technologies

Program Chairs: Panayiotis Zaphiris, Cyprus, and Andri Ioannou, Cyprus

Ruthi Aladjem, Israel
Abdulaziz Aldaej, UK
John M. Carroll, USA
Maka Eradze, Estonia
Mikhail Fominykh, Norway
Denis Gillet, Switzerland
Mustafa Murat Inceoglu, Turkey
Pernilla Josefsson, Sweden
Marie Joubert, UK
Sauli Kiviranta, Finland
Tomaž Klobučar, Slovenia
Elena Kyza, Cyprus
Maarten de Laat, The Netherlands
David Lamas, Estonia
Edmund Laugasson, Estonia
Ana Loureiro, Portugal
Katherine Maillet, France
Nadia Pantidi, UK
Antigoni Parmaxi, Cyprus
Borzoo Pourabdollahian, Italy
Janet C. Read, UK
Christophe Reffay, France
Nicos Souleles, Cyprus
Ana Luísa Torres, Portugal
Stefan Trausan-Matu, Romania
Aimilia Tzanavari, Cyprus
Johnny Yuen, Hong Kong
Carmen Zahn, Switzerland

External Reviewers

Ilia Adami, Greece
Iosif Klironomos, Greece
Maria Korozi, Greece
Vassilis Kouroumalis, Greece
Asterios Leonidis, Greece
George Margetis, Greece
Stavroula Ntoa, Greece
Nikolaos Partarakis, Greece

HCI International 2015

The 15th International Conference on Human–Computer Interaction, HCI International 2015, will be held jointly with the affiliated conferences in Los Angeles, CA, USA, in the Westin Bonaventure Hotel, August 2–7, 2015. 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://www.hcii2015.org/`

General Chair
Professor Constantine Stephanidis
University of Crete and ICS-FORTH
Heraklion, Crete, Greece
E-mail: `cs@ics.forth.gr`

Table of Contents

Mental Workload and Stress

Visual Perception

Cognitive Issues in Interaction and User Experience

Cognitive Psychology in Aviation and Space

Transport and Industrial Applications

Mental Workload and Stress

System Delay in Flight Simulators Impairs Performance and Increases Physiological Workload

Nina Flad[1], Frank M. Nieuwenhuizen[1], Heinrich H. Bülthoff[1,2], and Lewis L. Chuang[1,⋆]

[1] Department of Perception, Cognition and Action, Max Planck Institute for Biological Cybernetics, Tübingen
[2] Department of Cognitive and Brain Engineering, Korea University
{nina.flad,frank.nieuwenhuizen,heinrich.buelthoff, lewis.chuang}@tuebingen.mpg.de

Abstract. Delays between user input and the system's reaction in control tasks have been shown to have a detrimental effect on performance. This is often accompanied by increases in self-reported workload. In the current work, we sought to identify physiological measures that correlate with pilot workload in a conceptual aerial vehicle that suffered from varying time delays between control input and vehicle response. For this purpose, we measured the skin conductance and heart rate variability of 8 participants during flight maneuvers in a fixed-base simulator. Participants were instructed to land a vehicle while compensating for roll disturbances under different conditions of system delay. We found that control error and the self-reported workload increased with increasing time delay. Skin conductance and input behavior also reflect corresponding changes. Our results show that physiological measures are sufficiently robust for evaluating the adverse influence of system delays in a conceptual vehicle model.

1 Introduction

Delays between input and feedback in a closed-loop control task can result in both perceptual and control instabilities. For example, in head-slaved visualization systems (i.e., head-mounted virtual reality displays), temporal discrepancies between head movements and display updating can result in oscillopsia (also referred to as simulator sickness) in which the human operator perceives an unstable virtual environment that "swims around" his head [1]. In vehicle simulators, time delays between manual inputs and visual feedback can lead to notable increases in performance errors as well as perceived workload [2–4]. The latter can induce stress that induces physiological reactions in the autonomic nervous system.

⋆ The work in this paper was supported by the myCopter project, funded by the European Commission under the 7th Framework Program.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 3–11, 2014.

Previous studies of flight performance have shown that visual feedback delays can decrease performance and increase workload. For example, participants who performed a low-level flight task under visual lag conditions produced larger altitude errors and responded with higher workload scores on a questionnaire [5]. In a different study, increasing system delays decreased piloting performance as well as the subjective handling qualities of the aircraft, when pilots were required to perform a side-step maneuver in a helicopter simulation as well as actual test flight [4].

This decrease in performance can be more specifically attributed to the influence of visual feedback delays on closed-loop control performance. Trying to compensate for a time-delayed error has been shown to result in pilot-induced oscillations, wherein the control inputs from the pilot actually adds to the overall system disturbance instead of subtracting from it [2]. This is especially detrimental to the performance of precision tasks, such as hovering or landing. If the pilot is trained on such maneuvers in a simulator that suffers from time delays, more time is necessary to acquire the targeted skill and the transfer of training to real flight could be problematic, since a real aircraft with no system delays can be expected to respond differently [6]. Moreover, training with system delays has also been shown to increase the workload that is subjectively experienced by the pilot [5].

Many studies employ questionnaires to assess the participants' workload. Although this approach is well established, it has known weaknesses; the measurement is obtrusive and cannot be conducted during the task itself. Therefore, self-assessment of workload relies on the participant's recollection of the task, which could be subjectively altered.

An increase in workload can induce stress, which in turn leads to psychophysiological reactions of the autonomic nervous system. For example, induced workload can increase the heart rate as well as disrupt the regular fluctuations of inter-heartbeat-intervals [7]. In addition, it can widen the perspiratory glands and affect skin conductance [8]. Both heart activity and skin conductance can be measured using skin electrodes, thus providing an online metric for stress and workload during control activity itself.

In the current work, we investigate the influence of system delays on the control of a personal aerial vehicle (PAV) concept model [9]. We introduced delays of 0, 200, 400 or 600 ms and the influence of these delays was assessed in terms of our participants' control performance, control inputs, physiological responses and questionnaire results.

The remainder of this paper is organized as follows. Section 2 describes the flight task as well as the simulator and the aircraft model, followed by a description of data acquisition and analysis. Section 3 presents the results and possible interpretations. In section 4 we summarize our findings and discuss the implications for flight simulator studies.

Table 1. The parameters for the disturbance in the roll axis of the vehicle

i	a_i	ψ_i	f_i
1	0.5	0	0.0159
2	0.3	1	0.0796
3	0.2	0	0.0477
4	−0.2	1	0.0159

2 Methods

2.1 Participants

Eight male participants took part in our study. Their ages ranged between 22 and 34 years. All were researchers at the Max Planck Institute for Biological Cybernetics and had normal or corrected-to-normal vision.

2.2 Apparatus and Flight Model

We evaluated the effect of system delay of a dynamic PAV concept model in a fixed-base flight simulator. The simulator consisted of a display wall of nine screens taking a field-of-view of 105° by 100°. The participants used generic helicopter controls consisting of foot pedals, collective stick and cyclic stick.

The outside visualization was provided by Flightgear, an open-source flight simulator [10], while the control model was implemented in Matlab/Simulink and running at 256 Hz. The control model simulated the vehicle's dynamics and calculated the position and orientation of the aircraft based on the current control inputs. The outputs were then transformed into world coordinates and sent to the Flightgear computers that rendered the scene, namely San Francisco International Airport. The landing zone measured approximately 55 by 260 meters and was at the end of a runway (see Figure 1).

The PAV model represents an augmented helicopter with uncoupled cyclic, collective and pedal controls. Its rotational dynamics were of the Attitude Command-Attitude Hold (ACAH) type, such that a constant deflection of the cyclic stick resulted in a constant rotational attitude. Participants directly controlled a rate in the heave axis with the collective stick. A constant input on the pedals resulted in a specific rotational rate around the yaw axis. Subjects had full control over all the vehicle's degrees of freedom during each trial. In our experiment, a time delay of 0, 200, 400 or 600 msec was introduced between the control input and vehicle dynamics. These values were chosen based on a pilot study.

In addition, a disturbance was introduced in the roll axis during flight. Thus, our participants had to compensate for this disturbance even whilst performing the primary task of landing the PAV. The forcing function was a summation of four sinusoidal functions

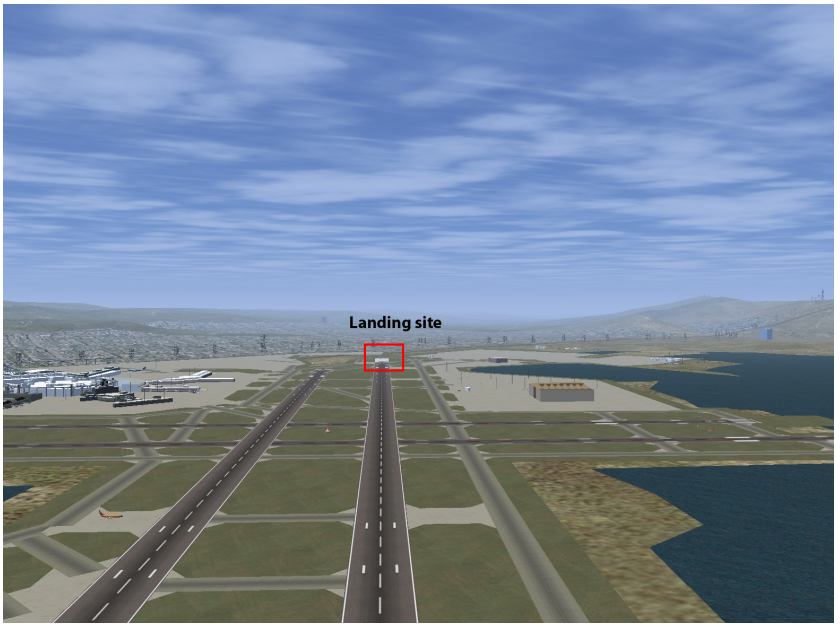


Fig. 1. Landing site as seen from the participants upon trial initiation

$$d(t) = \sum_{i=1}^{n} a_i * \sin(\psi_i * \frac{\pi}{2} + 2\pi * f_i * t) \tag{1}$$

with the parameters a_i, ψ_i and f_i given in Table 1.

2.3 Procedure

Our participants were instructed to fly the PAV from their initial position at the start of the trial towards the visible airport, follow the runway and land in a designated area at the end of the runway. In addition, they were required to maintain PAV stability and to compensate for disturbances in roll axis. Upon a successful landing, they were required to press a button to end the trial. Alternatively, each trial ended automatically if the maneuver was not successfully completed within eight minutes.

Prior to data collection, every participant had at least five one-hour training sessions with the simulator and the PAV model. During the first two training sessions, there were neither disturbance nor system delay to facilitate user familiarization with the control devices and the vehicle's dynamics. In addition, participants had to learn to navigate by relying on visual landmarks near the landing site. In the next two sessions, the roll disturbance was introduced, but without any time delay. Each training session always consisted of five flight sessions, separated by a thirty second break.

After the first four training sessions, the participants experienced at least one additional session under actual experimental conditions. The sessions for data collection consisted of four trials that varied in roll disturbance and time delay (0 ms, 200 ms, 400 ms and 600 ms) separated by a break of five minutes. In this break, participants were asked to rate workload with a digital version of the NASA-TLX questionnaire on a separate laptop. We collected ECG and skin conductance values during flight as well as during the breaks.

2.4 Data Collection and Analysis

Subjective workload was assessed using a computerized NASA Task Load Index (NASA-TLX). This rating scale consists of six independent scales, defined as follows [11]:

- Mental Demand (e.g. thinking, deciding, remembering, looking, etc.)
- Physical Demand (e.g. pushing, pulling, activating, etc.)
- Temporal Demand (e.g. time pressure)
- Performance
- Effort (required to achieve the level of performance)
- Frustration Level

This questionnaire was administered after every condition. It provides an overall workload score as well as scores for each individual scale and the composition of the overall score by the individual scales.

To measure performance, we calculated the root mean square error in compensating for the roll disturbances as well as control input activity. When participants experience a subjective loss of control or are performing badly, they tend to alter their input behavior or control effort. Therefore, we analyzed changes in the stick input activity. The stick input was collected at 256 Hz and analyzed in the frequency domain. We evaluated spectral densities in the frequencies higher than 0.1 Hz, since the disturbances took place in frequencies lower than 0.08 Hz. Thus, any changes in bandwidth above 0.1 Hz could be attributed to the pilot and not our disturbance function.

The first physiological measure is the skin conductance, which can be measured with a constant potential. The human skin possesses a natural electrical resistance, but contains sweat glands serving as conductive channels. Higher activity in the sweat glands results in lower resistance and better conductance [12]. The sweat glands are innervated by sympathetic activity only and, therefore, the skin conductance can serve as an indicator for stress and anxiety [8]. For the analysis, we normalized the mean conductance of each trial to a baseline measurement taken before the first test trial.

The second physiological measure is based on ECG measurements. The mean heart rate changes constant in response to changing environmental demands. These changes occur periodically and depend on the mental and physical state of the human. They are evoked by activity in the (para)sympathetic nervous system and have been found to be sensitive to work conditions, such as before and after

a driving task [7], or different phases of a monitoring and detection experiment [13]. We collected ECG data at 256 Hz and filtered it offline. The heart-beats in this signal were detected and the instantaneous heart-rate for every inter-beat-interval was calculated. We analyzed the spectral densities of the resampled time series in the 0.07–0.14 Hz band as well as the 0.15–0.4 Hz band. The power in the low frequencies is related to sympathetic activity, whereas the high frequency band is almost completely influenced by the parasympathetic nervous system in addition to respiration [14]. The low band is therefore widely regarded as the better measure for workload and stress.

3 Results and Discussion

All data was submitted to a one-way repeated measures analysis of variance (ANOVA) for the factor of time delay.

The questionnaire data showed an effect of system delay. The overall workload follows a linear trend ($F(3,21)=22.44$, $p<0.01$), indicating that increases in system delay induced higher subjective workload in our participants. The same linear trend can be found for the independent scales of the NASA-TLX. Interestingly, even though the self-rated performance decreased and the frustration increased, the participants did not "give up" on the task but increased their effort accordingly. In addition, the overall composition of the workload stayed the same (see figure 2). This suggests that the experimental manipulation of system delay increased subjective workload without changing the nature of the task itself.

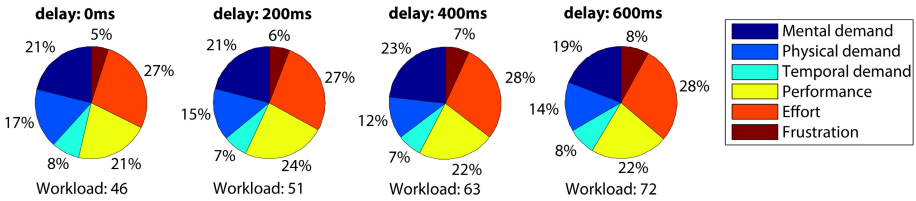


Fig. 2. Subjective workload increases with system delays but does not vary in its composition

As is the case with the self-rated performance, the objective error in compensating for the disturbance increased with increasing system delay ($F(3,21)=12.65$, $p<0.01$). Therefore, it follows that time delays have a deteriorating effect on the control task.

This deterioration in performance with increasing system delays evoked corrective inputs from the participants who tried to keep the vehicle stable. This is indicated by a linear increase in the power of the stick activity between 0.1 to 0.5 Hz ($F(3,21)=36.32$, $p<0.01$; see Figure 3). Since these disturbances take

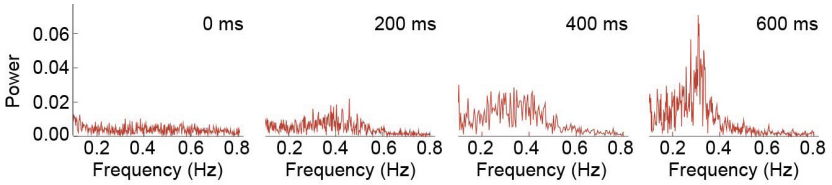


Fig. 3. Increasing input activity at frequencies higher than the disturbance. This can lead to pilot induced oscillations.

place at lower frequencies than these inputs, this behavior can destabilize the vehicle even more, resulting in pilot induced oscillations.

In correspondence with subjective workload measurements, an increase in system delays also resulted in higher skin conductance ($F(3,21)=5.72$, $p=0.01$, see figure 4). This indicates that participants experienced stress and arousal.

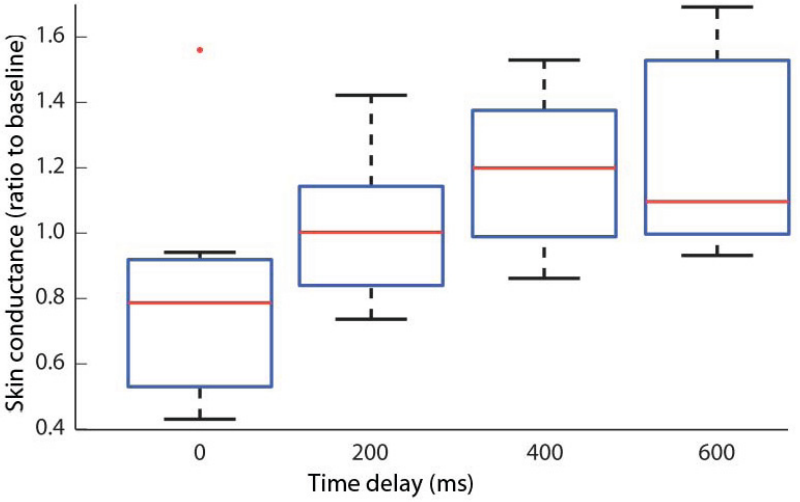


Fig. 4. The skin conductance increases linearly with increasing time delay

Nonetheless, the ECG measures for heart-rate and heart-rate variability did not show any significant changes to the manipulation of system delays. In addition, we did not find any changes between the test trials and breaks. Therefore, ECG based measures are not a reliable metric for stress and arousal in the current control task.

4 Conclusions

Overall, our findings show that delays between the control input and system response can impair control performance, elicit pilot-induced oscillations and increase workload, both in terms of self-reported and physiological measures. This is an important point to note in the design of virtual simulation systems, such as driving and flight simulators, that are intended for the purpose of training closed-loop control.

A system that is slow in responding to the human operator's input could induce the human operator to submit a larger response than is required for precise maneuvers. This results in a larger than intended vehicle response that needs to be corrected for subsequently. It could, thus, result in more errors than necessary and even instill counter-productive behavior that will have to be unlearned in the real world.

Our findings indicate that this loss of control has an impact on the operator's perceived and physiological workload. Therefore, system delays have a genuine influence on the operator's conscious sense of well-being as well as his physiological system.

In this work, we demonstrated that skin conductance activity can offer a complementary approach to the use of questionnaires. Changes in heart-based measurements might be too slow to indicate the changes in stress levels experienced by the human operator in our current experimental task. In contrast to a questionnaire, an unobtrusive measure such as this can be employed to assess multiple maneuvers in a complex mission. In addition to the traditional assessment of novel controller systems for their handling qualities, skin conductance measurements can allow the same systems to be evaluated for their physiological comfort.

To conclude, we demonstrate that system delays can detrimentally affect control performance due to pilot-induced oscillations. This has an adverse effect on the perceived workload of the operator as well as on his physiological system. The approach described here is a viable protocol for the evaluation of novel controller systems and simulators intended for closed-loop control.

References

1. Allison, R., Harris, L., Jenkin, M., Jasiobedzka, U., Zacher, J.: Tolerance of temporal delay in virtual environments. In: Proceedings IEEE Virtual Reality 2001, pp. 247–254 (2001)
2. Middendorf, M., Lusk, S., Whitley, J.: Power spectral analysis to investigate the effects of simulator time delay on flight control activity. In: AIAA Flight Simulation Technologies Conference, pp. 46–52 (1990)
3. Wildzunas, R.M., Barron, T.L., Wiley, R.W.: Visual display delay effects on pilot performance. Aviation, Space, and Environmental Medicine 67, 214–221 (1996)
4. Jennings, S., Reid, L.D., Craig, G., Kruk, R.V.: Time Delays In Visually Coupled Systems During Flight Test and Simulation. Journal of Aircraft 41, 1327–1335 (2004)

5. Middendorf, M., Fiorita, A., McMillan, G.: The effects of simulator transport delay on performance, workload, and control activity during low-level flight. In: AIAA Flight Simulation Technologies Conference (1991)
6. Riccio, G., Cress, J., Johnson, W.: The effects of simulator delays on the acquisition of flight control skills: Control of heading and altitude. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 1186–1290 (1987)
7. Zhao, C., Zhao, M., Liu, J., Zheng, C.: Electroencephalogram and electrocardiograph assessment of mental fatigue in a driving simulator. Accident; Analysis and Prevention 45, 83–90 (2012)
8. Chattopadhyay, P., Bond, A., Lader, M.: Characteristics of galvanic skin response in anxiety states. Journal of Psychiatric Research 12, 265–270 (1975)
9. Perfect, P., Jump, M., White, M.D.: Development of handling qualities requirements for a personal aerial vehicle. In: Proceedings of the 38th European Rotorcraft Forum, Amsterdam, Netherlands (2012)
10. Perry, A.R.: The Flightgear flight simulator. In: 2004 USENIX Annual Technical Conference, Boston, MA (2004)
11. Hart, S., Staveland, L.: NASA Task Load Index (TLX) v1. 0 users manual (1986)
12. Montagu, J., Coles, E.: Mechanism and measurement of the galvanic skin response. Psychological Bulletin 65, 261–279 (1966)
13. Tattersall, A., Hockey, G.: Level of operator control and changes in heart rate variability during simulated flight maintenance. The Journal of the Human Factors 37, 682–698 (1995)
14. Camm, A., Malik, M.: Heart Rate Variability: Standards of Measurement. European Heart Journal of the Physiological Interpretation and Clinical Use (1996)

Value Sensitive Design of Automated Workload Distribution Support for Traffic Control Teams

Maaike Harbers[1] and Mark A. Neerincx[1,2]

[1] Delft University of Technology, Delft, The Netherlands
[2] TNO Human Factors, Soesterberg, The Netherlands
M.Harbers@tudelft.nl, Mark.Neerincx@tno.nl

Abstract. This paper studies the effects of automated support for workload distribution in traffic control teams on human values such as security, autonomy and privacy. The paper describes a workshop in which the support system's stakeholders, their values, and the effects of the support system on these values were analyzed. The workshop results were used to derive design recommendations that minimize the negative effects on the stakeholders' values. The main conclusions are that in order to minimize negative impacts on privacy, trust and team spirit, the type and amount of information that is shared to improve workload distribution should be adjustable, depending on the role of the receiving party.

1 Introduction

Traffic management usually involves a team of operators that together control the traffic flow. One of the problems in traffic control teams is that workload is not always distributed evenly over team members [23]. A way to harmonize workload distribution in such teams is by adding an electronic partner (ePartner) [15] to the team that monitors the workload of the operators and shares this information with others. With this information, team members or the team leader can decide to take over or reallocate tasks, respectively. Thus, the ePartner provides automated support for workload distribution in traffic control teams.

Developing workload distribution support yields several challenges. The operators' activities and physical states need to be assessed, the ePartner should be able to reason about these observations, e.g. to determine workload and to decide when to inform whom, and the ePartner's interface should be usable. Besides these more technical and usage-oriented challenges, the design of an ePartner also yields ethical challenges, which we will focus on in this paper. An ethical issue, for instance, is how the ePartner affects the team members' values such as team spirit. Whereas the first type of challenges can be addressed with techniques from software engineering [24] and usability engineering [20], ethical challenges can be addressed by Value Sensitive Design [8].

Value Sensitive Design (VSD) is a methodology that aims to account for human values in the design of technology. The idea behind VSD is that humans esteem values such as autonomy, security, privacy, responsibility and well-being [6], and that technology can either support or hinder these values. In a

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 12–21, 2014.

human-automation team, for instance, automation should be *understandable* for its human team members, it should ensure their *safety*, and it should not hinder their *autonomy*. Hindering these values may lead to a decrease in motivation, which in turn can have a negative impact on team performance [4]. VSD offers a collection of theory, tools and methods to account for values in design.

In this paper, we describe how we used VSD techniques for the design of an ePartner that supports workload distribution in train traffic control teams. For that, we organized a Value Sensitive Design Workshop with subject matter experts in the domain of train traffic control. In the workshop, we presented an existing system for assessing someone's cognitive workload and emotional state [17], and sketched how this technology could be used to support workload distribution. Subsequently, we analyzed the stakeholders of this technology, their values, and the possible effects of different solutions on these values. Based on the outcome of the workshop, we derived a number of design recommendations for the ePartner to be developed.

The focus of this paper is not on a final solution, but on the design process and how we accounted for human values in that process. By that, the contribution of this work is two-fold. First, it contributes to the development of a value-sensitive ePartner supporting workload distribution for traffic control teams. Second, it contributes to the development of VSD by applying some of its techniques and discussing our experiences with that.

The outline of this paper is as follows. In section 2, we describe the train traffic control team we studied and how workload is currently distributed in that team, and we describe the solution we envision to improve workload distribution in this team. In section 3, we provide a short introduction into VSD. In section 4, we describe how we setup the VSD workshop and provide its results. In section 5, we discuss the implications of the workshop outcome for the design of the workload distribution support. In section 6, we end the paper with a discussion.

2 Workload in Train Traffic Control Teams

For this research, we studied train traffic control teams in ProRail, the organization that is responsible for controlling train traffic in The Netherlands. The Dutch railway network is used by multiple passenger and cargo transporters. All this train traffic is managed from thirteen regional control centers and one national control center. The regional control centers are occupied by a team of train traffic controllers (treindienstleiders) and a team leader. In this section we first describe how workload is currently distributed over time and team members in the train traffic control teams we studied. Subsequently, we describe the CLES monitor and ePartner by which we aim to improve the distribution of workload in these teams.

2.1 Current Situation

In the traffic control teams we studied, the distribution of workload over time is rather uneven. In a normal situation, train traffic is automatically regulated by a

train traffic control system that follows fixed schedules describing the train traffic flow. When there is a small disruption, the system automatically reschedules the trains. Under these circumstances, train traffic controllers mainly monitor the situation, and their workload is rather low. In case of a larger disruption, however, the system cannot provide satisfying solutions, and is usually switched off by the train traffic controllers. Instead of relying on the system, they manually regulate train traffic by controlling the signals and switches on the rails, and informing train drivers about the changes. Depending on the complexity of the disruption, this may yield high levels of workload.

The distribution of workload over team members is based on the division of railway sections over train traffic controllers. Each train traffic controller is responsible for a particular section of the railway network, e.g. one station, and performs the work associated to that section. Thus, when a disruption only affects certain sections of the network, it may occur that the train traffic controllers responsible for those sections experience high levels of workload, whereas the others do not have much to do. It is possible that train traffic controllers perform tasks that are associated to railway sections of colleagues to alleviate their workload. In practice, however, this rarely happens because train traffic controllers do not always know how much workload their colleagues experience. Moreover, they tend to solve their own problems as much as possible and ask for help at a relatively late stage of a disruption.

2.2 Adding Automation

In the train traffic control domain, distribution of workload over time can hardly be controlled due to the unpredictability of disruptions on the railway network, but the distribution of workload over team members can be changed. To improve workload distribution in a team, insight in the distribution of workload over members is needed.

Motived by the current, at times uneven distribution of workload in operational teams, Neerincx et al. [17] developed a CLES monitor that assesses workload of train traffic controllers by measuring their Cognitive Load and their Emotional State. The cognitive load measurement is an implementation of Neerincx's model for cognitive task load [14, 18], describing the effects of task allocations on performance of operators working in dynamic, critical and high-demand task environments. The emotional state measurement assesses the operators' arousal and valence, which predict the affective load they experience [11, 16]. With these two measures an operator's workload can be determined. For details about the CLES monitor we refer to [17].

Collecting information with a CLES monitor in itself does not lead to an improved workload distribution. We envision a solution in which an ePartner, representing the operator, reasons about the information gathered by the CLES monitor, and informs (the ePartners of) other team members when necessary. When an ePartner receives information from another ePartner, it can provide (part of) this information to the person it represents, e.g. by showing the information on an awareness or observability display [1–3].

This solution requires policies that describe *what* information an ePartner should share *when* and *with whom*. There are many possibilities. Information can involve actual cognitive load and emotional state values, a value that combines both values, an indication of an unusual pattern, or a suggestion to take over someone's work. The information can be shared continuously, every n seconds, only if the train traffic controller gives permission, in emergency cases, or when the CLES values are high or unusual. Information can be shared with only to the train traffic controller being monitored, to his team leader, to all members of the team, or only to team members with a low workload. Based on these options, policies for the ePartner can be, for instance, that it is obliged to continuously provide all information about the operator it represents to the team leader, or that it is authorized to provide cognitive load information to team members with a low workload if the operator gives permission.

3 Value Sensitive Design

In this section we provide a short introduction to Value Sensitive Design (VSD), the methodology we use to account for values in the design of the ePartner. VSD has been developed over the last 25 years and can be defined as follows [8].

> *Value Sensitive Design is a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process.*

Central to VSD are analyses of direct and indirect stakeholders, and their values. A value refers to what a person or group of people considers important in life. Values that could play a role in the design of technology are, for instance, autonomy, security, privacy, safety, trust, responsibility, sustainability, and fun.

Designers are confronted with value tensions. For example, supporting the value of security, e.g. by placing more surveillance cameras, may hinder privacy, and supporting safety by making actions with a safety risk impossible, may hinder a user's autonomy. VSD aims to make designers aware of these tensions during the design process so that they make informed design choices. The objective is to strive for improvements rather than perfection.

VSD includes investigations on three levels: the conceptual, empirical and technological level. Conceptual investigations involve analyses of direct and indirect stakeholders, and how their values are affected by technology. Empirical investigations can be interviews, surveys, and prototype testing. Technological investigations refer to both the examination of existing technological solutions and the development of new technology. These investigations are intended to be iterative, so that the designer can modify the design continuously.

This paper describes our conceptual investigations regarding the ePartner providing automated workload distribution support. Within VSD, a collection of techniques for conceptual investigations has been developed, such as value scenarios [13], value dams and flows [12], and envisioning cards [7]. The workshop we held is most similar to the value dams and flows method. The value

dams and flows method starts with identifying stakeholders of the (envisioned) system, then for each stakeholder group the potential harms and benefits of the system are identified, and then the values underlying those harms and benefits are identified. The workshop we held also started with the identification of stakeholders. The second step of our workshop, however, involved the identification of values of these stakeholders, and only after that, the effects of the system on these values were identified. The reason for changing the second and third step of the value dams and flows method was to bring even more focus on values rather than on the system design.

4 VSD Workshop

4.1 Setup

The workshop was held at the railway company ProRail with four participants who were all employed by ProRail. Three of the participants were ICT experts responsible for the design and development of new ICT systems, and one of them was a research coordinator. At the start of the workshop, none of the participants was familiar with the VSD methodology. The workshop took two hours and was lead by the first author of this paper, assisted by a colleague who recorded the whole session on video.

The workshop started with an interactive presentation about the role of values in the design of technology to introduce the participants to the basic ideas of VSD. Subsequently, the CLES monitor and the envisioned solution of the ePartner supporting workload distribution were presented to the participants. Several possible displays to provide information were shown, such as a continuous graph of cognitive load and emotional state values, and display with messages reporting about abnormal cognitive load and emotional state values. Then, the participants were asked to perform a stakeholder and value analysis regarding the CLES monitor. Following the workshop setup we described in the previous section, the workshop participants discussed the following questions.

1. Who are direct and indirect **stakeholders** of the monitor?
2. What are their **values**, relevant with respect to the monitor?
3. What are the **effects** (positive or negative) of the monitor on these values? Give an explanation for this effect.

The findings were written down on a white board. The workshop ended with a short discussion about the value of VSD for developing human-automation teams for the railway organization.

4.2 Results

The participants identified two direct and nine indirect stakeholders. The direct stakeholders they identified were: train traffic controller (TTC) being monitored, and team leader. The indirect stakeholders they identified were: TTC in operation, TTC in preparation or evaluation, regional traffic controller, train driver,

Stakeholder	Value	Effect	Explanation of the effect
TTC being monitored	Insight	+	Monitor gives a TTC more insight in his own functioning
	Openness	+	Monitor ensures that a TTC provides more openness towards others
	Trust	-	Monitor can be interpreted as if a TTC is not trusted in letting others know about the height of his workload
	Privacy	-	Monitor discloses possibly privacy sensitive information about a TTC
	Recognition	+/-	Hardworking TTCs can feel recognized when the monitor shows others how hard they work, but it can also be seen as a sign of distrust in how hard they are working, i.e. they are not being recognized
	Power	-	Monitor makes a TTC more vulnerable, e.g. when others see that he performs poorly
Team leader	Trust	+	Monitor increases a team leader's trust in his own judgments when the monitor confirms his intuitions about TTCs
	Insight	+	Monitor gives a team leader more insight in the functioning of TTCs
	Support decisions	+	Monitor can confirm a team leader's assumptions about the functioning of TTCs on which he bases his decisions
	Control	+	More insight increases a team leader's feeling of being 'in control'
Team member, also TTC	Being important	+	Taking over work of a TTC with high workload gives a team member the feeling that he is useful and important
	Politeness	+	Monitor ensures that when a TTC has a high workload, team members will only disturb him with really important matters
	Helpfulness	+	A team member knows better when to help a TTC when he has insight in his workload
	Team spirit	+/-	Helping each other when it is busy increases team spirit, but constantly monitoring each other's functioning decreases team spirit
	Curiosity	+	Monitor satisfies curiosity in each other's functioning
	Availability	+	Monitor ensures that a team member is more quickly available if that is really necessary

Fig. 1. Stakeholders, values and effects identified in the VSD workshop

transporters, travelers, broadcast, TTCs in next shift, and leader safety operation. Figure 4.2 displays the identified values of the most important stakeholders, and the effects of the CLES monitor on these values.

The stakeholders 'TTC in operation' and 'TTC in preparation or evaluation' were identified to have the same values, effects and explanations. In the table they are therefore depicted as 'team member, also TTC'. Interestingly, TTCs in operation and TTCs in preparation or evaluation were initially identified as indirect stakeholders, whereas team leaders were identified as direct stakeholders. In the value and effect analyses, however, this distinction seemed to disappear, as in the explanations, all were envisioned to be able to receive workload information about team members. It could therefore be argued to consider all of these as direct stakeholders.

The results show that, according to the workshop participants, the monitor supports most of the stakeholders' values. In their view, the values trust, privacy and power were hindered by the monitor, and the values recognition and team spirit could be both supported and hindered by the monitor.

5 Implications for Design

We aim to use the results of the value and stakeholder analyses for the design of the ePartner. However, neither the value dams and flows method nor any other VSD technique offers much guidance on how to incorporate the results of a conceptual analysis in an actual design [10, 5]. For instance, there is no structured method for what to do in case two values are in conflict with each other. Furthermore, VSD offers no concrete method for translating stakeholder and value analyses into actual design requirements. In this section, we will therefore adopt a rather informal approach to derive design recommendations from the workshop results.

As mentioned earlier, three important questions in designing workload distribution support are: what information should be shared, when, and with whom? Examining the participants' explanations in Figure 4.2, they all seem to concern the situation where most or all information gathered by the CLES monitor is shared to all other participants. Therefore, in this section we will explore how we can minimize the negative effects of the CLES monitor by sharing less information, while maintaining the positive effects associated to sharing information. This section is organized according to the three stakeholder groups, starting with the TTC being monitored, followed by team members and the team leader.

An examination of the values of the **TTC being monitored** shows that only the value 'insight' refers to sharing the information collected by the CLES monitor to the TTC being monitored himself. The associated explanation is that 'the monitor gives a TTC more insight in his own functioning'. For a maximum insight in one's functioning, it is helpful to receive information about actual values of cognitive load and emotional state continuously. Providing such information to the TTC being monitored does not hinder any of the other values, neither of the TTC, nor or the others. The only danger can be that when CLES values are displayed for the TTC, that others may see them on the TTC's display as well, which would go against the TTC's values of privacy and power. To overcome this problem, the TTC should have the freedom to switch off the display. Moreover, for situations where a TTC switches off the display or when he is too busy to look at the display, he should have the possibility to inspect the data at a later moment.

For **team members**, receiving information about a colleague TTC has mostly positive effects. For the TTC being monitored, however, that has mixed effects. On the one hand, sharing information with others supports the value of openness, but on the other hand, it might hinder the values of trust, privacy and power. The explanations show that most of the team members' values relate to helping team members when necessary, either by taking over tasks or not disturbing them. To support these values, it is sufficient to receive suggestions about whom to help every now and then, and continuous information about a TTC's cognitive load and emotional state is not necessary. This option removes the negative effect on team spirit, as team members cannot constantly monitor each other's functioning this way, and it greatly relieves the negative impacts on the TTC's values of trust, privacy and power. Though this solution has large benefits, it

may decreases the positive effects of the monitor on 'curiosity' of team members, and on 'openness' of the TTC being monitored.

It was indicated that for the **team leader** receiving information about a TTC has only positive effects. The reasons for these positive effects are related to the team leader's desire to have insight in the functioning of the TTC. To support these values, it is likely that team leaders would like to receive more rather than less information. For the TTC being monitored, however, this will hinder his values of trust, privacy, recognition, and power. Thus, there is a value tension for which no easy solution is available. We suggest to search for a middle ground in this situation, e.g. by only sharing workload information every now and then (not continuously), only in case of a disruption in the environment, or only if the TTC agrees on doing that.

To summarize the above discussion into concrete recommendations for the design of an ePartner that harmonizes workload distribution, we believe that the ePartner should be adaptive with regard to what information should be shared with whom. First, the ePartner should provide cognitive load and emotional state values to the TTC being monitored. Second, the ePartner should limit the information it shares with team members to letting them know when their help is needed, and when they should not disturb the TTC being monitored. Third, the ePartner should be able to provide different types of information to the team leader, e.g. depending on the TTC's preferences, the team leader's preferences, and the situation at hand. Following these recommendations should minimize the negative effects of the monitor on the stakeholders' values.

6 Discussion

The work presented in this paper is part of a project that aims to develop automated support for improved workload distribution in train traffic control teams, while taking human values into account. In this paper we described our view of an ePartner that collects information about the workload of a TTC, and shares this information with (ePartners of) the TTC's team members and his team leader. The intended result of this information sharing is that TTCs more often take over each other's tasks and thus establish a harmonized workload distribution. To account for human values in the design process of such an ePartner, we analyzed its stakeholders, their values, and its effects on these values in a VSD workshop. The results showed that monitoring and sharing TTCs' cognitive load and emotional state supports most stakeholder values, but that it may hinder the TTCs' values of privacy, trust and team spirit. In the previous section, we made several recommendations to overcome these hindrances.

The VSD workshop we held was an adaptation of the value dams and flows method in the sense that in our workshop values were identified before the effects of the technology on values, rather than the other way around. According to our observations, the adaptation was successful and workshop participants had no difficulties in coming up with values. The reactions of the workshop participants on the workshop were positive. In a final discussion about the use of VSD for

developing human-machine interfaces for train traffic controllers, participants indicated that VSD offers a valuable perspective that often receives too little attention in their organization. Participants stated that it would be valuable to organize a similar VSD workshop with direct users of the envisioned technology, i.e. with train traffic controllers and team leaders. Finally, the results of VSD workshop are in line with other findings in the literature, in which tensions between information sharing and privacy were encountered as well [22, 21].

As mentioned in Section 5, VSD offers no structured method for moving from value analyses to an actual design. This could be seen as a shortcoming of VSD [10], but it may also be that this is not part of VSD's objectives. Despite the lack of a concrete method, we were able to derive design recommendations. For future work, however, we believe that it would be beneficial to develop a more structured method. This would make the process more transparent and it could enhance the quality of the recommendations that are derived. We suggest an approach that draws on techniques from situated Cognitive Engineering [19] to derive and refine requirements specifications, and techniques from requirements engineering [9] to prioritize over multiple requirements.

In future work, we will implement a prototype of the ePartner following the recommendations derived in this paper. This is in line with the VSD methodology, prescribing rapid prototyping and an integrative approach of conceptual, empirical and technical investigations. After implementing a first prototype, we plan to test our design in a user study, and investigate whether workload distribution is actually improved and whether the support system supports the stakeholders' values.

Acknowledgments. This research was conducted within the RAILROAD project and is supported by ProRail and the Netherlands organization for scientific research (NWO) (under grant 438-12-306). We would like to thank all participant of the VSD workshop for their valuable contributions, and Alex Kayal for his assistance during the workshop.

References

1. Carroll, J.M., Rosson, M.B., Convertino, G., Ganoe, C.H.: Awareness and teamwork in computer-supported collaborations. Interacting with Computers 18(1), 21–46 (2006)
2. Kraut, R., Dabbish, L.: Awareness displays and social motivation for coordinating communication. Information Systems Research 19(2), 221–238 (2008)
3. de Greef, T., van der Kleij, R., Brons, L., Brinkman, W.-P., Neerincx, M.: Observability displays in multi-teams. In: NDM (2011)
4. DeChurch, L.A., Mesmer-Magnus, J.R.: The cognitive underpinnings of effective teamwork: a meta-analysis. Journal of Applied Psychology 95(1), 32 (2010)
5. Detweiler, C., Hindriks, K., Jonker, C.: Principles for value-sensitive agent-oriented software engineering. In: Weyns, D., Gleizes, M.-P. (eds.) AOSE 2010. LNCS, vol. 6788, pp. 1–16. Springer, Heidelberg (2011)

6. Flanagan, M., Howe, D.C., Nissenbaum, H.: Embodying Values in Technology: Theory and Practice, pp. 322–353. Cambridge University Press (2008)
7. Friedman, B., Hendry, D.: The envisioning cards: A toolkit for catalyzing humanistic and technical imaginations. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems, pp. 1145–1148. ACM (2012)
8. Friedman, B., Kahn, P.H., Borning, A.: Value sensitive design and information systems. In: Human-Computer Interaction and Management Information Systems: Foundations, pp. 348–372 (2006)
9. van Lamsweerde, A.: Requirements Engineering. John Wiley & Sons (2007)
10. Manders-Huits, N.: What values in design? The challenge of incorporating moral values into design. Science and Engineering Ethics 17(2), 271–287 (2011)
11. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology 14(4), 261–292 (1996)
12. Miller, J.K., Friedman, B., Jancke, G.: Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. In: Proceedings of the 2007 International ACM Conference on Supporting Group Work, pp. 281–290. ACM (2007)
13. Nathan, L.P., Klasnja, P.V., Friedman, B.: Value scenarios: a technique for envisioning systemic effects of new technologies. In: CHI 2007 Extended Abstracts on Human Factors in Computing Systems, pp. 2585–2590. ACM (2007)
14. Neerincx, M.A.: Cognitive task load design: model, methods and examples, pp. 283–305. Lawrence Erlbaum Associates, Mahwah (2003)
15. Neerincx, M.A., Grant, T.: Evolution of electronic partners: Human-automation operations and epartners during planetary missions. Journal of Cosmology 12, 3825–3833 (2010)
16. Neerincx, M.A.: Modelling cognitive and affective load for the design of human-machine collaboration. In: Harris, D. (ed.) HCII 2007 and EPCE 2007. LNCS (LNAI), vol. 4562, pp. 568–574. Springer, Heidelberg (2007)
17. Neerincx, M.A., Harbers, M., Lim, D., Van der Tas, V.: Automatic feedback on cognitive load and emotional state of traffic controllers. In: The Current Issue (2014)
18. Neerincx, M.A., Kennedie, S., Grootjen, M., Grootjen, F.: Modeling the cognitive task load and performance of naval operators. In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) FAC 2009. LNCS, vol. 5638, pp. 260–269. Springer, Heidelberg (2009)
19. Neerincx, M.A., Lindenberg, J.: Situated cognitive engineering for complex task environments. In: Naturalistic Decision Making and Macrocognition, pp. 373–390 (2008)
20. Nielsen, J.: Usability engineering. Elsevier (1994)
21. Nissenbaum, H.: Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press (2009)
22. Olson, J.S., Grudin, J., Horvitz, E.: A study of preferences for sharing and privacy. In: CHI, pp. 1985–1988. ACM (2005)
23. Porter, C.O., Hollenbeck, J.R., Ilgen, D.R., Ellis, A.P., West, B.J., Moon, H.: Backing up behaviors in teams: the role of personality and legitimacy of need. Journal of Applied Psychology 88(3), 391 (2003)
24. Pressman, R.S., Ince, D.: Software engineering: a practitioner's approach, vol. 5. McGraw-Hill, New York (1992)

Transparency of Automated Combat Classification

Tove Helldin[1], Ulrika Ohlander[2], Göran Falkman[1], and Maria Riveiro[1]

[1] University of Skövde, Sweden
firstname.lastname@his.se
[2] Saab Aeronautics, Sweden
Firstname.Lastname@saabgroup.com

Abstract. We present an empirical study where the effects of three levels of system transparency of an automated target classification aid on fighter pilots' performance and initial trust in the system were evaluated. The levels of transparency consisted of (1) only presenting text–based information regarding the specific object (without any automated support), (2) accompanying the text-based information with an automatically generated object class suggestion and (3) adding the incorporated sensor values with associated (uncertain) historic values in graphical form. The results show that the pilots needed more time to make a classification decision when being provided with display condition 2 and 3 than display condition 1. However, the number of correct classifications and the operators' trust ratings were the highest when using display condition 3. No difference in the pilots' decision confidence was found, yet slightly higher workload was reported when using display condition 3. The questionnaire results report on the pilots' general opinion that an automatic classification aid would help them make better and more confident decisions faster, having trained with the system for a longer period.

Keywords: Classification support, automation transparency, uncertainty visualization, fighter pilots.

1 Introduction

Fighter pilots must be able to discriminate between various categories of aircraft during a mission. They must further be able to prioritize amongst the detected targets in order to ensure mission efficiency. However, due to factors, such as stress, data over-load and fast-paced decision-making situations, it is not always possible for the pilots to perform their classification and prioritization tasks with good quality. As stated by de Jong et al. in [8], about 15% of the military defense engagements are against friendly targets and severe accidents have occurred (see for example [2,20]). Several causes have been listed as contributing to the engagements of non-hostile targets, such as inexperienced military personnel, insufficient data quality, failure of the battle management systems, classification criteria, rules of engagement and malfunction of the identification system [2,4].

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 22–33, 2014.

To decrease the error rates associated with target identification and classification, efforts have been made to improve the threat evaluation support systems used, and guidelines for information visualization and operator-automation interaction have been proposed (see for instance [11,6]). However, there is a lack of empirical studies where the effects of applying all or some of these guidelines on expert operators' trust and performance have been evaluated. This paper addresses this gap by investigating the effects of applying three levels of automation transparency on expert fighter pilots' performance and trust when classifying targets incorporated into scenarios developed by domain experts.

The paper is structured as follows: Section 2 summarizes related work, such as how the task of target classification is conducted as well as what it meant by system transparency. Section 3 describes the proof-of-concept target classification prototype implemented, whereas section 4 presents the study performed. Section 5 summarizes the results obtained, and sections 6 and 7 provide discussions and conclusions and ideas for future work, respectively.

2 Related Work

2.1 Target Classification

The class of a target reveals target features such as allegiance, intent and possible capabilities. The class can be assessed through analyzing sensor data (such as analyzing identification-friend-foe (IFF) replies according to pre-defined database setups), kinematical data (such as g-force, speed and altitude which reveal target behavior and platform performance characteristics) and through investigating team-based information (such as if another team-member has classified the target) (see [5,9] for more information).

Different target classes can fulfill certain combinations of attributes of technical and behavioral characteristics and capabilities. For example, a fighter aircraft has a typical radar signature, may be able to fly at high speeds and can fulfill certain attack profiles [9]. These attributes can be measured by different sensors, and through matching these attributes with a database collection of known attribute setups, a probable target class can be generated. Perfect matches are not always possible, due to non-complete databases, errors in sensor readings as well as due to countermeasures used by the adversary, masking the targets' characteristics.

To perform their classification tasks, fighter pilots have to quickly analyze incoming data. When the identity or class of a detected target is unknown, the pilots have to act fast to redirect sensors, analyze the data and collaborate within the military teams to be able to create a better base for establishing the class of the object. However, due to the fast-paced decision-making tempo, the large amounts of data provided and the severity of the tasks they perform, fighter pilots might not always be able to make correct decisions.

2.2 System Transparency

In various domains, support systems are used that are able to aid operators to, for example, collect and analyze information, but also to generate recommendations of actions, as well as the implementation of these actions. However, as stated by Paradis et al. [16], it is not likely that a human operator will accept an automatically generated decision or action if no explanation is provided regarding how the system arrived at this conclusion, which might lead to misuse and disuse of the automated aid [10,17]. Further, if important information regarding the system performance and inferences is omitted from the primary system displays, negative effects such as flawed operator decision-making and accidents can occur. Therefore, several researchers have argued for the importance of *system transparency*, highlighting the need for the operators to be able to easily use and understand how a support system works [13,18]. The majority of research related to improving system transparency is concerned with the visualization of additional meta-information, or information qualifiers, especially the visualization of uncertainty (see e.g. [1,3,12,19]). However, there are some works that mark the need for presenting the reliability associated with system generated recommendations (see e.g. [14,15,21]). For example, in Wang et al. [21], it was concluded that the presentation of the reliability of an automated target identification system improved the participants' performance. However, the participants in the study performed by Wang et al. were not expert operators (students and employees at the Department of National Defense). Further, the participants in the study were not informed of the classification model of the system, something which might have affected the participants' trust and performance.

3 The Classification Prototype

To evaluate the effects of automating the classification analysis tasks on the fighter pilots' performance and initial trust in an automated target classification system, a proof-of-concept target classification prototype was implemented. The classification model uses rules as basis for the evaluations, where setups of rules and associated threshold values are used to determine if a target belongs to the class "fighter" aircraft, "attack" aircraft or "other". A "fighter" aircraft is defined as a target that can possibly carry a beyond visual range missile, an "attack" aircraft as a fixed-wing aircraft with attack capacity, whereas an aircraft of class "other" is defined as neither a fighter nor attack aircraft. The rules incorporate parameters such as the maximum speed, maximum altitude, maximum g-force, type of engine and the identity extracted from the electronic warfare system, and pre-defined threshold values for the numerical data and membership rules for the categorical data are used to classify targets into one of the three categories. For example, a target with altitude > 50000 ft indicates a fighter aircraft, a target with altitude > 35000 ft can either belong to the class fighter or attacker, whereas a target with a altitude < 35000 ft can belong to all the three classes. Results from the different parameter measurements are fused and the most probable class is extracted. If inconsistent parameter checking results are obtained, i.e.

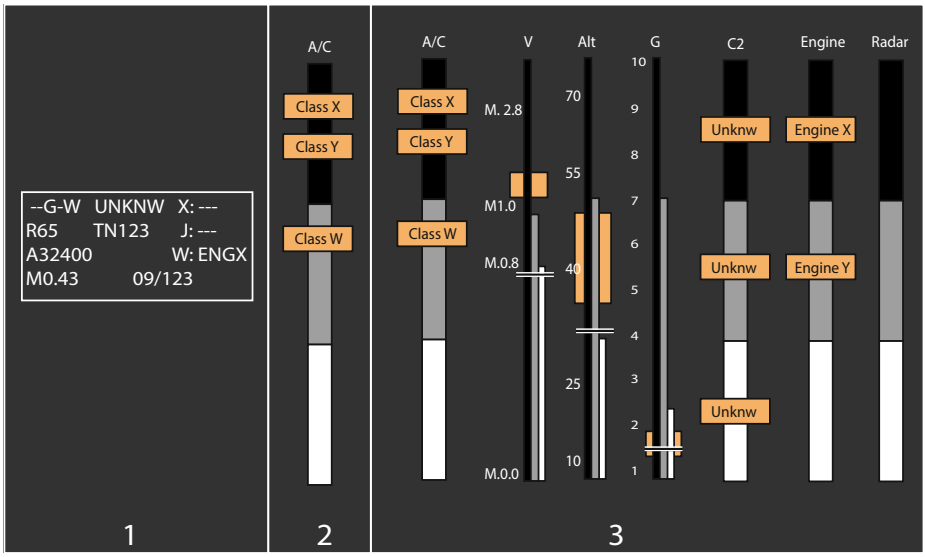


Fig. 1. The display conditions representing different levels of class automation transparency. Figure section 1 represents the first display condition, including a text-based representation of the object parameters. The second display condition, figure section 2, includes both the text-based parameters and a system generated class suggestion. The third display condition, figure section 3, includes the text-based parameters, the system class suggestion as well as a detailed view of the included parameters with associated (uncertain) historical sensor values.

that the target can either be a fighter or attacker for example, both conclusions are displayed.

The results from the classification inferences are displayed through three means, representing three levels of classification transparency: (1) only presenting text-based information regarding the specific object (without any automated support), (2) accompanying the text-based information with an automatically generated object class suggestion and (3) adding the incorporated sensor values with associated (uncertain) historic values in graphical form (see Figure 1). These display conditions are hereafter referred to as display condition (DC) DC1/DC2/DC3. In DC2 and DC3, the results from the classification inferences are shown with a representation based on intervals. This representation was chosen due to the possibility of making the classification model transparent to the operators, i.e. by making the parameter rule threshold values explicit to the pilots.

Three colors were used to indicate the different classification categories, i.e. targets positioned within the black segment of the interval representation belong to the "fighter" class, whereas targets in the gray and white segments belong to the classes "attacker" or "other" respectively. Uncertainties in the parameter inputs are indicated through two means: dealing with numerical data, the larger

the orange bar, the greater the uncertainty, while dealing with categorical data, the term 'unknown' is used. If no value for a parameter has been measured, the associated interval bar is left blank. The measured maximum value for a numerical parameter is indicated through the use of a black and white line.

By using the classification prototype described above, we wanted to investigate the effects of automating the classification analysis tasks on the fighter pilots' performance and initial trust in the automated target classification system. We hypothesized that:

- H1: The pilots will need more time to classify a target when being presented with DC2 and DC3, than DC1.
- H2: The pilots will report higher confidence ratings when being presented with DC3, than DC1 and DC2.
- H3: The pilots will be better able to make correct decisions when being presented with DC2 and DC3, than DC1.
- H4: The pilots will report higher trust ratings when being presented with DC3, than DC1 and DC2.

4 Method

4.1 Participants

Six experienced fighter pilots participated in the study. The pilots were all male. The average age was 51.8, and the pilots had on average 2700 hours of experience of flying military aircraft.

4.2 Experimental Design and Procedure

The classification rules were applied to a set of 11 scenarios using a fixed arrangement of targets and associated characteristics. Each scenario was further divided into 3 sub-scenarios, each containing one of the three display conditions. Altogether, 33 momentary images were prepared by domain experts and presented to the participants. The pilots were to classify one target per scenario, i.e. 33 targets in total. The order of the targets to classify was randomly generated using a balanced Latin square design. The target classification prototype was shown on a 24-in. LCD monitor (set to 1920x1200 pixel resolution). A regular desktop station with a computer screen, a mouse and a keyboard was used during the experiments.

A briefing was held before the test session to inform the pilots of the purpose of the experiment, the classification model used and the scenario used during the experiment (i.e. the possible targets to appear in the scenarios). The pilots were also provided time to study the classification rules used and were trained to understand the different visualizations by going through two training scenarios. During the test session, the pilots were presented with the different scenario images and were requested to make a classification decision based on each of them. To their aid, they were provided with the "classification system manual"

in paper form, i.e. the rules of the system prototype, the possible target types to appear during the scenarios etc. After the test session, the pilots were further asked to answer a questionnaire.

4.3 Collected Data

The time needed for the pilots to make a decision regarding a specific target was noted. When having classified a target (i.e. stating that the target was a fighter, attacker or other), the pilots were asked which parameters they based their classification decision on. The pilots were further asked how certain they were about their decisions being correct, as well as their perceived workload (on a scale between 1–3, where 1 indicates a low certainty/low workload). Moreover, in the scenarios where the system provided a classification suggestion, the pilots were asked to estimate their level of trust in the automatically generated class suggestion (on a scale between 1–3, where 1 indicates low trust). After the test session, the pilots answered a questionnaire, containing questions regarding their perceived workload when using the three display conditions, as well as their general subjective opinions of the classification support system (see the Appendix).

5 Results

Due to the different nature of the data analyzed, various statistical tests have been carried out. Paired t-tests (for normal distributed data and more than two variables), Mann-Whitney U (for categorical data) and Kruskal-Wallis (for discrete or not normal distributed data) tests with significance level $\alpha = 0.05$ were conducted over the collected data, using the SPSS statistical software.

The test results show that the average time needed to make a classification increases with the level of information presented, i.e., the largest classification time corresponds to DC3, while the lowest was recorded using DC1 (DC1: mean = 9.275, std. deviation = 8.882; DC2: mean = 10.378, std. deviation = 6.063; DC3 : mean = 12.898, std. deviation = 7.627). There was a significant effect regarding the time needed to classify the targets between DC2 and DC3 ($t(65) = -2.799$; $p = 0.007$) and between DC1 and DC3 ($t(65) = -3.270$; $p = 0.002$), see Figure 2.

The average trust value assigned by the participants to DC3 was higher (mean = 2.924, std. deviation = 0.703) than the one assigned to DC2 (mean = 1.909, std. deviation = 0.738). A t-test shows that there is a significant difference in trust values between these two displays ($t(65) = -4.749$; $p =< 0.000$).

The number of correct classifications was very similar for the three conditions, and no significant differences were found during the analysis of the means (DC1: mean = 4.000, std. deviation = 1.732; DC2: mean = 4.545, std. deviation = 1.213; DC3: mean = 4.454, std. deviation = 1.809). Neither was a significant difference found between the different display conditions regarding the pilots' confidence values (DC1: mean = 1.969, std. deviation = 0.743; DC2: mean = 1.954, std. deviation = 0.773; DC3: mean = 2.121, std. deviation = 0.832).

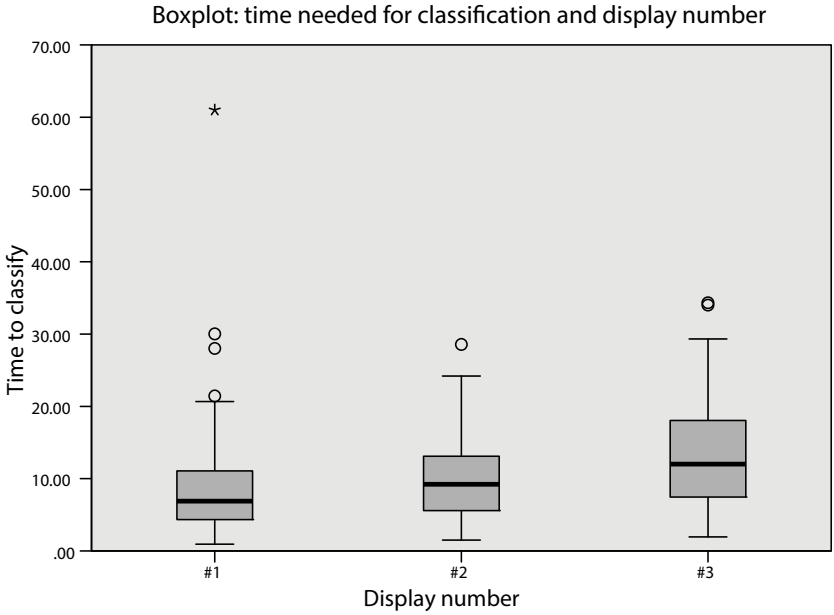


Fig. 2. Time to classify – the pilots needed more time to make a classification decision when being presented with DC2 and DC3, than when using DC1. Further, the pilots needed more time when using DC3 than DC2.

The trust in automation scale provided by Jian et al. [7] was adapted to provide an indication of the pilots' initial trust in the classification support system prototype used. The questionnaire template includes questions such as how well respondents understand the workings of the automation and if they would like to use the system (see the Appendix for the full set of questions), answered by using a 7-point Likert scale. All six pilots marked a value between 5–7 on the question *"I understand how the classification support system works"*, i.e. indicating a good understanding of the system. Four pilots marked 5 or 6 regarding if they would use the system if it was provided to them, indicating a good acceptance of the system, and 5 pilots argued that the system would have a positive effect on their performance in classification scenarios (marking 5 or 6). In summary, the responses to the trust questionnaire indicate that the majority of the pilots indeed trusted the system and that they would use it if it was provided to them.

The pilots also answered questions regarding their subjective impressions of the three display conditions, if the classification suggestion automatically provided to them (if applicable) influenced their decision, as well as if the uncertainty representations affected their decisions. The pilots indicated that DC1 and DC3 were most useful for them since the design provided a visualization of all the parameters used in the evaluation. They further argued that when

parameter values were missing, the historical values provided when using DC3 (i.e. the maximum speed/altitude etc.), were helpful. However, due to to their previous experience with DC1 and its limited amount of information included, two of the pilots argued that they felt that they made a quicker decision during this condition. However, five of the pilots argued that the information provided through DC3 provided them with a greater information base for grounding their decisions, possibly leading to more accurate decisions. All six pilots argued that if a classification suggestion was provided to them, it influenced their own classification decision—either through making them analyze the different parameters once more if they disagreed with the system generated suggestion, or through confirming their thoughts, making them feel more confident in their decisions. Only two pilots indicated that the uncertainty visualizations affected their decisions, possibly due to the simplicity of the scenarios. The questions in the questionnaire further indicate that the pilots experienced slightly higher workload when using DC3.

6 Discussion

In summary, a statistical significant difference regarding the time needed to make a classification decision was noted between DC2 and DC3 ($p = 0.007$), and between DC1 and DC3 ($p = 0.002$), thus confirming our first hypothesis. This difference was also noted in the questionnaire results where two pilots explicitly noted that they felt that they needed more time to make a decision when using DC3 than display DC1. This result is perhaps not surprising given that the amount of information incorporated into DC3 exceeds the information given by DC1. Further, several pilots argued that they are more accustomed with the representation found in DC1, thus making it easier for them to extract the information needed to make a decision.

A significant difference was also found regarding the pilots' trust in the classification support system between DC2 and DC3 ($p < 0.000$). Answers from the questionnaire highlight the fact that many of the pilots found DC2 too much "black box" for them, i.e. masking the underlying parameters used in the evaluations, resulting in decreased understanding and trust. No statistical difference between DC1 and DC3 was found, thus not completely fulfilling our fourth hypothesis, i.e. that the pilots would report higher trust values when using DC3. However, different trust ratings might be obtained if further evaluations and training with the system using DC3 are conducted. Yet, it might also be the case that the pilots viewed the setup of the system rules too simplistic to reflect appropriate classification bases to be used during real missions, perhaps lowering their trust in the classification system suggestions as a whole.

No statistical significance regarding the pilots' confidence in their decisions or the correctness of these decisions was noted in the quantitative data. However, results from the data collected from the questionnaire questions indicate that the pilots found that the historical values provided in DC3 indeed positively affected their decisions, i.e. making them more confident in the decisions made.

Thus more research is needed to investigate our second hypothesis (i.e. that DC3 will generate higher confidence ratings) after the pilots have trained with the system for a longer period, which will probably also have an effect on the pilots' subjective trust ratings toward the support system.

7 Conclusions and Future Work

The results indicate that the pilots needed more time to make a classification decision when being provided with DC2 and DC3 than display condition DC1 (H1) (significant difference, see Figure 2). Comparing the reported trust ratings from DC2 and DC3, higher ratings can be found in DC3 (H4) (significant difference). This provides and indication that the inclusion of the different parameter values in the classification visualization provides a better foundation for pilot trust in the classification system, yet at the expense of longer decision times. However, no trust difference was found between DC1 and DC3. No differences in decision confidence or decision correctness were found when comparing the three display conditions (H2, H3). The answers to the questionnaire questions highlight the slight increase of perceived workload, but also the general opinion that such support system would aid them make better and more confident decisions faster, having trained with the system for a longer period.

Future work will include further evaluations of different levels of transparency of the automatic classification system after conducting longer training sessions together with the pilots. Future work could also include investigations of other visualization formats where the amount of information elements used during DC3 could be reduced, perhaps resulting in a decrease of the pilots' feelings of information overload when being presented with DC3. Such information element reduction could result in greater performance improvements when using the proposed classification support system.

Acknowledgment. This research has been supported by VINNOVA (Swedish Governmental Agency for Innovation Systems) through the National Aviation Engineering Research Program (NFFP5-2009-01315), Saab AB, the University of Skövde. We would like to thank the study participants, Johan Holmberg, Jens Alfredson and Marike Brunberg for their valuable feedback and support.

References

1. Bisantz, A.M., Cao, D., Jenkins, M., Pennathur, P.R., Farry, M., Roth, E., Potter, S.S., Pfautz, J.: Comparing uncertainty visualizations for a dynamic decision-making task. Journal of Cognitive Engineering and Decision Making 5(3), 277–293 (2011)
2. British Ministry of Defence: Military Aircraft Accident Summary: Aircraft accident to Royal Air Force Tornado GR MK4A ZG710. Tech. rep. (2004)
3. Dong, X., Hayes, C.: Uncertainty visualizations helping decision makers become more aware of uncertainty and its implications. Journal of Cognitive Engineering and Decision Making 6(1), 30–56 (2012)

4. Fisher, C., Kingma, B.: Criticality of data quality as exemplified in two disasters. Information & Management 39(2), 109–116 (2001)
5. Gelsema, S.: The desirability of a nato-central database for non-cooperative target recognition of aircraft. In: Proceedings of the RTO SET Symposium on Target Identification and Recognition Using RF Systems, Oslo, Norway, October 11-13 (2004)
6. Irandoust, H., Benaskeur, A., Kabanza, F., Bellefeuille, P.: A mixed-initiative advisory system for threat evaluation. In: Proceedings of the 15th International Command and Control Research and Technology Symposium: The Evolution of C2, Santa Monica, California, USA (2010)
7. Jian, J.Y., Bisantz, A., Drury, C.: Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics 4(1), 53–71 (2000)
8. de Jong, J., Burghouts, G., Hiemstra, H., te Marvelde, A., van Norden, W., Schutte, K.: Hold your fire!: Preventing fratricide in the dismounted soldier domain. In: Proceedings of the 13th International Command and Control Research and Technology Symposium: C2 for Complex Endeavours, Bellevue, WA, USA (2008)
9. Krüger, M., Kratzke, N.: Monitoring of reliability in bayesian identification. In: 12th International Conference on Information Fusion, FUSION 2009, pp. 1241–1248. IEEE (2009)
10. Lee, J., See, K.: Trust in automation: Designing for appropriate reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society 46(1), 50–80 (2004)
11. Liebhaber, M., Feher, B.: Air threat assessment: Research, model, and display guidelines. In: The Proceedings of the 2002 Command and Control Research and Technology Symposium (2002)
12. MacEachren, A.M., Roth, R.E., O'Brien, J., Li, B., Swingley, D., Gahegan, M.: Visual semiotics & uncertainty visualization: An empirical study. IEEE Transactions on Visualization and Computer Graphics 18(12), 2496–2505 (2012)
13. Mark, G., Kobsa, A.: The effects of collaboration and system transparency on cive usage: an empirical study and model. Presence: Teleoperators & Virtual Environments 14(1), 60–80 (2005)
14. McGuirl, J., Sarter, N.: Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Human Factors: The Journal of the Human Factors and Ergonomics Society 48(4), 656–665 (2006)
15. Neyedli, H., Hollands, J., Jamieson, G.: Beyond identity incorporating system reliability information into an automated combat identification system. Human Factors: The Journal of the Human Factors and Ergonomics Society 53(4), 338–355 (2011)
16. Paradis, S., Benaskeur, A., Oxenham, M., Cutler, P.: Threat evaluation and weapons allocation in network-centric warfare. In: 8th International Conference on Information Fusion, vol. 2, pp. 1078–1085. IEEE (2005)
17. Parasuraman, R., Sheridan, T., Wickens, C.: A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 30(3), 286–297 (2000)
18. Preece, J., Rogers, Y., Sharp, H.: Interaction Design: Beyond Human-Computer Interaction. Wiley, New York (2002)

19. Skeels, M., Lee, B., Smith, G., Robertson, G.G.: Revealing uncertainty for information visualization. Information Visualization 9(1), 70–81 (2010)
20. Smith, C., Johnston, J., Paris, C.: Decision support for air warfare: Detection of deceptive threats. Group Decision and Negotiation 13(2), 129–148 (2004)
21. Wang, L., Jamieson, G., Hollands, J.: Trust and reliance on an automated combat identification system. Human Factors: The Journal of the Human Factors and Ergonomics Society 51(3), 281–291 (2009)

Appendix

Questionnaire Template – Trust in the Classification Support System

The following questions were used to estimate the fighter pilots' initial trust in the classification support system used during the experiment. The questions were answered using a 7-point Likert scale (1: not at all, 7: to a high degree) The questions were adapted from the trust in automation questionnaire provided by Jian et al [7].

- I understand how the classification support system works – its goals, actions and output.
- I would use the classification support system if it was available to me.
- I believe that the classification support system will have a positive effect on my performance when it comes to classifying objects (faster decisions, more correct decisions).
- I put my trust in the system.
- I believe that the classification support system would aid me perform my classification tasks in a safe manner.
- I believe that the classification support system is reliable.
- I believe that I can trust the classification support system.

Questionnaire Template – Mental Workload, Situation Awareness and Trust

The following questions were posed to the fighter pilots to collect their subjective opinions of the usefulness of the classification support system, as well as their opinions of the three display conditions on their perceived mental workload, situation awareness and trust. Both Likert scale questions (from 1: not at all, to 10: highly agree) and free-text based questions were used.

- Would a classification support system aid you when performing your classification tasks during a mission? (Free text.)
- Did you have sufficient knowledge of the classification model to use the system? (Free text.)
- With which alternative (1, 2 or 3) do you feel you performed the best? Why? (Free text.)
- Estimate your mental workload when using visualization alternative 1 (Likert scale.)

- Estimate your mental workload when using visualization alternative 2 (Likert scale.)
- Estimate your mental workload when using visualization alternative 3 (Likert scale.)
- How well do you think that visualization alternative 1 supported your situation awareness? (Likert scale.)
- How well do you think that visualization alternative 2 supported your situation awareness? (Likert scale.)
- How well do you think that visualization alternative 3 supported your situation awareness? (Likert scale.)
- How easy was visualization alternative 1 to use? (Likert scale.)
- How easy was visualization alternative 2 to use? (Likert scale.)
- How easy was visualization alternative 3 to use? (Likert scale.)
- Do you think that visualization alternative 1 involves a risk to make a wrong classification decision? (Likert scale.)
- Do you think that visualization alternative 2 involves a risk to make a wrong classification decision? (Likert scale.)
- Do you think that visualization alternative 3 involves a risk to make a wrong classification decision? (Likert scale.)
- To which degree do you trust the information presented when using visualization alternative 1? (Likert scale.)
- To which degree do you trust the information presented when using visualization alternative 2? (Likert scale.)
- To which degree do you trust the information presented when using visualization alternative 3? (Likert scale.)
- When the classification support system provided a class suggestion, did you trust this suggestion? (Free text.)
- When the classification support system provided a class suggestion, did this suggestion affect your own decision? (Free text.)
- Did the uncertainty representation affect your decision-making? (Free text.)
- Did the uncertainty representation gave you a better understanding of the automatic classification? (Free text.)
- Did the uncertainty representation affect your trust in the system classifications? (Free text.)
- Would you like to manually establish rules and threshold values for the automatic classification? (Free text.)

System Requirements for an Advanced Cockpit to Reduce Workload and Stress

Paul M. Liston* and Nick McDonald

Centre for Innovative Human Systems, School of Psychology,
Trinity College, University of Dublin, Ireland
{pliston,nmcdonld}@tcd.ie

Abstract. This paper describes the requirements elicitation process and the subsequent system requirements for an advanced cockpit to reduce crew workload and stress. The paper outlines the need for a step-change in technology and operational practices to ensure the continued safety of a transport system which is predicted to grow. The ACROSS project aims to develop advanced cockpit solutions to reduce workload and stress in an increasingly congested aviation transport system. Six types of requirements were derived including aviate requirements, and navigate, communicate, manage systems, crew monitoring, and crew incapacitation requirements. The research project is currently specifying the human factors requirements for the technologies to achieve improved operational safety.

Keywords: requirements, flight deck, flight crew, cockpit, workload, stress.

1 Introduction

Aviation has achieved an enviable reputation as the leading transport sector in terms of safety. Decades of research and operational innovation have contributed to successive reductions in accident rates. Though there is a decreasing trend in the rate of fatal accidents in the period from 2002 to 2011 to 0.6 per million flights flown (CAA, 2013) the forecasted growth in air traffic of 4.7% annually (Airbus, 2013) means that aviation accidents will continue to occur on a regular basis.

52% of the fatal accidents from 2002 to 2011 involved a flight-crew related primary causal factor. Indeed, seven of the top-ten causal factors of all fatal accidents in this period came from the flight crew. The most frequently allocated causal factors were "Flight-handling" and "Omission of action or inappropriate action" (CAA, 2013).

1.1 The Challenge

As technology has progressed aviation accidents based solely on technical faults or problems have decreased. Indeed it is an entirely natural trajectory for accident

* "The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. ACP2-GA-2012-314501". ACROSS Project Website: www.across-fp7.eu

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 34–41, 2014.

statistics to increasingly implicate the human factor as technology develops. In the case of flight operations the humans in the cockpit are the factors introducing variability of performance. However, separating out the technology from the human operator – in this case the pilots – may not be useful in helping to understand why accidents and incidents occur. One possible reason for the high rate of flight-crew related accidents could be the increased complexity of aircraft systems. "Flight Crew Perception and Decision-making" (i.e. omission of action and inappropriate action) is the number one causal factor allocated for all fatal accidents for the ten-year period 2002 to 2011 (UK CAA). Research has shown that automation in the cockpit has led to issues relating to safety and situational awareness in particular (Mosier et al., 2013; Dao et al., 2009).

The predicted growth in air transport brings about further challenges. Not only are aircraft becoming more complicated, but the airspace and airports are becoming increasingly crowded (something compounded when weather conditions become sub-optimal). Couple this with the fact that the flight phases that have a high workload (approach, landing, turn-around and take-off) are happening in, and around, airports and it is clear that pilot stress and peak workload are key safety challenges for the aviation sector.

1.2 The Need

A step-change in technology and flight operations practices is needed to meet the aforementioned challenges. This step-change should not supplant ACARE SRA2's goal of increasing safety by a factor of 5 compared to the situation in 2000, but should supplement it and ensure that humans and machines work together effectively in the aviation system. Airframers and equipment suppliers have already focused their attention on reducing the complexity of aircraft technology and some of these solutions to improve cockpit operations have been implemented in new generation aircraft such as the Boeing B787 and the Airbus A380.

Despite these efforts certain combinations of unpredictable situations, such as difficult meteorological conditions, multiple system failures or cockpit crew incapacitation, can lead to peak workload conditions. These situations are difficult to anticipate and the number of actions that flight crew have to simultaneously execute and the amount of information they have to process can quickly render the workload unacceptably high. Given that accidents are more likely to occur when flight-crew workload is high, improving crew performance in peak workload conditions is thus critical to enhance safety. Clearly a more integrated, comprehensive solution is needed to address not just complexity, but also peak workload and stress.

1.3 The Solution

The achievement of an overall reduction in the number of aviation accidents necessitates the design and implementation of new solutions (based on both hardware and software) to allow flight crew to more easily manage peak workload situations. To this end, the ACROSS research project aims to make important safety gains by

developing an advanced cockpit to reduce workload and stress. Ensuring safe operations under crew peak workload can be achieved by providing a cockpit environment that mitigates the risk of human error on the flight deck and by limiting crew workload to ensure that pilots will have the opportunity to address all relevant issues in a timely and appropriate manner - thereby mitigating the risk of pilot error. Flight crew incapacitation is one circumstance in which the remaining pilot has to manage the situation under significant stress. In 2004 pilot incapacitation occurred on 36 occasions (Evans & Radcliffe, 2012). Pilot incapacitation can come about for various reasons and the ACROSS project intends to develop technologies which will help the remaining crew to manage these unplanned situations of reduced crew and to ensure the safe completion of the flight.

1.4 ACROSS Objectives

The ACROSS project has three main objectives which inform all research and development activities. They are:

1. Facilitate the management of peak crew workload situations during a flight
2. Allow reduced crew operations
3. Identify open issues for possible single-pilot operations

ACROSS Objective 1: Crew under peak workload situations
This objective targets fully capacitated crew with peak workload and will develop and demonstrate solutions up to Technology Readiness Level (TRL) 5 (Component and/or mock-up validation in a relevant environment).

ACROSS Objective 2: Reduced crew operations
This objective targets three conditions of reduced crew and will develop and demonstrate solutions up to TRL 3 (Analytical and experimental critical functions and/or characteristics proof-of-concept);

1. Long haul flight with reduced crew;
2. One pilot incapacitated;
3. Short/medium range flight with both pilots incapacitated.

ACROSS Objective 3: Open issues for possible single-pilot operation
Single-pilot operations in all conditions are considered a long-term evolution that is not in the scope of the ACROSS project. The project consortium considers single-pilot operations as a case study that stimulates innovation and facilitates the identification of solutions that could be used to improve the current safety level in situations of peak workload and reduced crew. Any solutions developed to manage peak workload and reduced crew situations may be considered for possible single pilot operations in the future.

2 ACROSS Requirements

The first step in realising the three ACROSS objectives was to elicit requirements for the systems which can help improve crew performance in peak workload and stressful conditions. 12 stakeholder organisations were involved in generating and specifying the requirements (these included aircraft manufacturers, aerospace safety service providers, navigational information providers, communication, crew and fleet management services providers, and cockpit communication services providers). Each stakeholder organisation was responsible for producing requirements according to their area of expertise.

2.1 ACROSS Requirements Process

Step 1 – Selecting the Scenario
The ACROSS project identified 29 scenarios based on accidents and current safety threats that were selected as relevant due to their implications for stress, workload and incapacitation on the flight deck. As such these scenarios were the starting point for the requirements elicitation process.

Step 2 – Using the Scenario as a reference
When the requirement authors were writing the requirement they were instructed to keep the specifics of the scenario in mind. The use of scenarios to guide the requirements process creates a constant link from the overall goals of the project right the way through evaluation and validation.

Step 3 – Completing the "Requirements Capture Form"
A 'Requirements Capture Form' was developed in order to gather the information about not just the system requirement but also the operations, processes and systems implicated. In addition the form structure ensured the requirements were linked to scenarios and objectives. Guidelines on how to write a good requirement (Kar & Bailey, 1996) were made available to the partners.

Step 4 – Collating and analysing the requirements
All requirements received were reviewed for completeness and clarity.

Step 5 – Requirement Stakeholder Clarification Interview
Teleconference interviews (guided by a standardised interview schedule) were held with the requirement authors. The objective was to refine and/or clarify any ambiguities or errors in the requirement text and to understand completely the background to the requirement – ensuring all sections of the 'Requirement Capture Form' were completed. Additional information about processes affected by the requirement, together with predicted impact were elicited.

Step 6 – External Experts Review
External experts (especially representing flight operations and ATM (air traffic management)) were presented with a selection of the collated requirements in order to solicit feedback from stakeholder perspectives not present in the requirement author group. The objective was to get feedback and suggestions for improvement from experts external to the project (and in particular those with operational experience).

Step 7 – Gap Analysis
A gap analysis was performed - the objective of which was to ensure that all crucial areas of interest had related requirements. All project stakeholders participated in the gap analysis – reviewing both the content of individual requirements and the overall scope of the requirements in their entirety. Requirement authors had the opportunity to accept or reject (citing reasons) any comments or suggestions. This process was managed using a Gap Analysis Protocol form and another one for Gap Analysis Resolution.

Step 8 – Fill Gaps Identified and Review
This was the last step in the requirements capture process. Any suggestions which were accepted at the Gap Analysis stage were actioned and reviewed before the requirements could be considered final.

3 Results – The Requirements

139 Requirement Capture Forms were elicited in Step 3 of the process. At the end of the gap analysis process – following the combination of requirements due to repetition and the specification of new requirements - there were 123 final requirements (see Table 1 for their distribution across the project's technical functions).

Table 1. Final ACROSS Requirements by Technical Function

Technical Function	Number of Final Requirements
Aviate	11
Navigate	27
Communicate	18
Manage Systems	8
Crew monitoring	10
Crew incapacitation	49

A selection of the final requirements are listed below (italicised).

Aviate
Aviate requirements relate to the task of flying the aircraft according to operational requirements, supported by monitoring from the pilot not flying.

- *The system shall assist the crew in performing manual tasks during the execution of initial climb.*
- *The system shall assist the crew when having to perform an emergency descent towards nearest, suitable airport in difficult terrain.*

Navigate
Navigate requirements relate to monitoring threats to the flight plan from weather, traffic, loss of infrastructure capability, and adapting the flight plan if necessary.

- *The system shall provide more functional and less physical references regarding aircraft status (mask the complexity of the system)*
- *The system shall offer the capability for the ground to support the remaining pilot to handle the situation.*

Communicate
Communicate requirements relate to maintaining contact between the crew and ATC (Air Traffic Control), cabin staff (if applicable), and the AOC (Airline Operations Centre).

- *The system shall maintain air-ground communication without requiring actions by the cockpit crew or ground crew.*
- *The system shall be robust against intentional interference (i.e. jamming) and non-intentional interference.*

Manage Systems
Manage systems requirements relate to monitoring, evaluating (and reconfiguring if necessary) the aircraft's systems status to ensure optimum efficiency and safety.

- *The cockpit and all its systems shall provide the relevant information through a functional view regarding aircraft remaining resources/performance, especially in case of abnormal or unexpected events.*
- *The system shall prioritize and filter EICAS / ECAM messages (Engine Indication and Crew Alerting System / Electronic Centralized Aircraft Monitor).*

Crew-monitoring
Crew-monitoring requirements relate to providing crew monitoring functions for the evaluation of the crew's physiological and behavioural condition as they operate the systems, and to adequately address peak workload situations and reduced crew operations.

- *The system shall be able to detect clues of vigilance loss, fatigue or peak workload.*
- *The system shall make the crew aware of their physiological and behavioural condition without being physically or psychologically intrusive.*

Crew-incapacitation

Crew-monitoring requirements relate to the extreme situation of incapacitated crew, which is encountered very rarely in a two-pilot configuration, and involve identifying and developing automatic functions and safety nets.

- *The system shall be capable of ensuring the continued safe flight and landing of the aircraft without any flight crew intervention after the start of take-off in normal conditions.*
- *The system shall automatically detect when one or more of the flight crew are incapacitated.*

4 Design and Development Implications

Eliciting requirements from scenarios is a recognised practice in requirements engineering. The approach outlined in this paper supplemented this practice with early consultation with a wide group of end-users beyond just the technology development team and the stakeholders represented in the project consortium.

Looking at the number of requirements elicited per technical function one could make an inference about the relative complexity of one proposed technical solution over another. Indeed 'Crew Incapacitation' requirements number 49, while the next function in order of number of requirements elicited is 'Navigate', with 27 requirements. Clearly the technology needed to resolve a situation where both pilots are incapacitated represents a step-change and is certainly cutting-edge but it doesn't mean that it is necessarily a more demanding proposition than supporting the navigation of an aircraft in peak workload conditions just because of the number of requirements elicited. At the outset of the requirements elicitation process the stakeholders were encouraged to define what was meant by "the system" in each requirement. In the clarification interviews it became clear that some partners were choosing to not define "the system", not to introduce ambiguity, but to build in a level of flexibility. In so doing no particular solution is precluded *a priori*. Given that a single requirement can impact more than one project objective or technical function it was considered prudent to have this level of flexibility at the requirements stage. Later, once the development work is underway the requirements can be specified in terms of what is meant, in that context, by "the system". Some of the requirements which were derived are at a 'high level' and others are more detailed and technical. At this initial stage this is not necessarily a limitation as in the 'Requirement Capture Form' all details relating to the requirement are captured – especially details about which objective is linked to each requirement. In this way those requirements that are overarching (and as such relate to all objectives) will feature in all targeted solutions. Those requirements which are very detailed and specific will only be passed on to those solutions that are related to them.

The next step for this project involves the specification of human factors requirements related to the system requirements detailed herein – how do the systems and processes that constitute the aviation transport system as it currently exists need to change in order to support the new technologies and solutions that will be developed in the ACROSS project? Parts of the information required to answer this question were elicited as part of the clarification interview (specifically information related to operational practices and processes) and the next step in the research involves gathering this information in a more systematic way so that the technology development work can be informed by a human factors-led agenda.

References

1. Airbus, Global Market Forecast (2013), http://www.airbus.com/company/market/forecast/
2. CAA. CAP 1036 - Global Fatal Accident Review 2002 to 2011, UK CAA (2013)
3. Dao, A.-Q.V., Brandt, S.L., Battiste, V., Vu, K.-P.L., Strybel, T., Johnson, W.W.: The Impact of Automation Assisted Aircraft Separation on Situation Awareness. In: Salvendy, G., Smith, M.J. (eds.) Human Interface, Part II, HCII 2009. LNCS, vol. 5618, pp. 738–747. Springer, Heidelberg (2009), http://human-factors.arc.nasa.gov/publications/Dao_et_al_ImpactofAuto_HCII09.pdf
4. Evans, S., Radcliffe, S.A.: The annual incapacitation rate of commercial pilots. Aviat. Space Environ. Med. 83, 42–49 (2012)
5. Kar, P., Bailey, M.: Characteristics of Good Requirements. International Council of Systems Engineers, Requirements Working Group. INCOSE Symposium (1996)
6. Mosier, K.L., Ute Fischer, U., Morrow, D., Feigh, K.M., Durso, F.T., Sullivan, K., Pop, V.: Automation, Task and Context Features Impacts on Pilots' Judgments on Human-Automation Interaction. Journal of Cognitive Engineering and Decision Making 7, 377–399 (2013), doi:10.1177/1555343413487178

Automatic Feedback on Cognitive Load and Emotional State of Traffic Controllers

Mark A. Neerincx[1,2], Maaike Harbers[1], Dustin Lim[1], and Veerle van der Tas[1]

[1] Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands
[2] TNO, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
mark.neerincx@tno.nl, M.Harbers@tudelft.nl

Abstract. Workload research in command, information and process-control centers, resulted in a modular and formal Cognitive Load and Emotional State (CLES) model with transparent and easy-to-modify classification and assessment techniques. The model distinguishes three representation and analysis layers with an increasing level of abstraction, focusing on respectively the sensing, modeling, and reasoning. Fuzzy logic and its (membership) rules are generated to map a set of values to a cognitive and emotional state (modeling), and to detect surprises of anomalies (reasoning). The models and algorithms allow humans to remain in the loop of workload assessments and distributions, an important resilience requirement of human-automation teams. By detecting unexpected changes (surprises and anomalies) and the corresponding cognition-emotion-performance dependencies, the CLES monitor is expected to improve team's responsiveness to new situations.

Keywords: resilience engineering, workload, affective computing, electronic partners, traffic management.

1 Introduction

Train traffic management has to deal with a complex internal and external environment, in which all kind of disturbances can appear with possibly cascading effects. Automation changed, and will further change, the operational control & supervision processes [1]. Automation has become an actor in these processes. Human and automation have to deal jointly with possible conflicts and dependencies between solutions, means and resources during the disturbance responses. Two types of research questions can be distinguished, centering (1) on the actual process of dealing with anomalies or (2) on the realization of a resilient human-automation team. For the first type, an example research question for train traffic control, is "how to improve distributed situation awareness and balance workload across operators through sharing of information on (i) what remote others are doing, (ii) how they are progressing with their tasks, and (iii) how their task progression affects the common goal". For the second type, an example research question for train control is "how to improve the competencies and skills of operators through situated feedback on (i) their physical and mental condition, (ii) their relationships with team-members, and (iii) their work attitude and motivation".

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 42–49, 2014.

A knowledge-based approach is required to answer these questions and to develop corresponding methods for achieving resilience, e.g., by providing support on performance, behavior, intention, task progression, and mental and physical condition of (remote) actors to increase coordination among actors and to subsequently allocate resources flexibly [2]. Such support improves the capabilities to mutually empower the human and automation for disturbance anticipation, monitoring, responding, and learning. For such mutual empowerment, we aim at ePartners that help humans in complex dynamic environments, and as such improve resilience [3, 4]. Such ePartners should have knowledge of the (momentary) capacities of the traffic controller and the current task and context in which she or he operates, and be able to assess the fit of these capacities to the task demands. To enhance resilience, the ePartners and human operators should be able to share, complement and correct each other's assessment of the workload distribution among humans and automatons.

There is a rich history and enormous amount of research on workload in the human factors domain (e.g., [5]) and on emotion in the affective computing domain (e.g., [6]). However, workload models, techniques and applications are diverse with different levels of granularity. Moreover, there is a lack of models that both (i) are understandable and accessible for the operators themselves, and (ii) formalize the interrelationships between the cognitive and affective processes for realistic complex settings. In other words, for real-world traffic control settings, there is not yet a sound, self-explaining formal model to automatically assess and guide cognitive and affective processes coherently. This paper presents the incremental development of such a model, to be implemented in a support tool for real-time balancing of cognitive and affective load.

2 Background

For several years, we conducted research to derive a transparent, coherent and concise workload model from established human factors theories and empirical studies in realistic process control settings. Taking an incremental development approach, we aimed at a modular model that can be fed or instantiated with information from both "machine and human sensors" [7]. The modules should allow for human-machine sharing of knowledge, for job design, mission planning and dynamic (real-time) workload allocation. This research was conducted in the train traffic control, naval ship center control, and space domain.

First, it was concluded that classical workload analyses for train traffic control mainly assessed the ***Time Occupied (TO)*** of train dispatchers, disregarding the *cognitive* demands or the work. To address these demands, a ***Cognitive Task Load (CTL)*** analysis was developed that assesses the required ***Level of Information Processing (LIP)*** (among others, based on the Knowledge, Rule and Skill-framework of Rasmussen, [8]). Simple and routine tasks evoke more efficient (i.e., a lower level of) information processing, whereas complex and new tasks require more extensive and intensive processing. The new CTL-analysis method proved to provide better assessments than the "old" method: the identification of inadequate workload

distribution in control posts, the reveal of context-dependencies and the proposal of a standard [9].

Subsequently, a third CTL-factor was added to the model, ***Task-Set Switches (TSS)***, to address the work demands of process control operations that may have to respond to unforeseen events or alarms immediately [10]. These demands entail both the responses on a single event after a long period of indifference and the responses on a large number of, almost co-occurring, alarms. The 3-dimensional model identifies specific types of CTL-states (properties): overload, under-load, vigilance, cognitive lock-up and optimal load [11]. The corresponding CTL-analysis method provided adequate predictions of the task load and identified negative effects on operator performance of under- and overload situations in a naval ship control center [12]. Furthermore, applying a "Naïve Bayesian network" for predicting performance from the CTL-values, provided performance estimations with 86% and 74% accuracy for a, respectively, high-fidelity simulator and real ship control center [13].

However, this CTL-model does not address the affective processes of work, which have a major impact on the performance in our high-demand application domains. Therefore, an Emotional State (ES) model was being constructed that complements the CTL-model. The combined model is called the Cognitive Load and Emotional State (CLES) model.

3 The CLES Model

In this section we explain the internal model of the CLES tool. After providing a general overview, we give more detailed discussions on the modeling of cognitive load and emotional state, respectively. We will pay attention in particular to the use of fuzzy logic for emotional state modeling, enabling to reason (i) in a transparent way (ii) with variables that have a truth value between 0 and 1 [14].

3.1 Three Layers of CL and ES Analysis

The monitor consists of three layers in which cognitive load and emotional state are represented and analyzed. The layers have an increasing level of abstraction. The activities performed in each layers are the following.

1. Sensing of cognitive load and emotional state
2. Modeling of cognitive load and emotional state
3. Reasoning about cognitive load and emotional state

In the first layer, information about the operator being monitored is perceived. More specifically, the CLES tool perceives observables regarding the operator's physical state and the tasks he performs. In the second layer, these observables are used to model cognitive load and emotional state. This results in a cognitive load value and an emotion classification, respectively. In the third layer, the CLES tool reasons about the cognitive load and emotional state of the operator, e.g. to detect surprising changes or rare combinations.

3.2 Cognitive Load Modeling

Cognitive load modeling in the CLES tool is based on Neerincx et al.'s model of cognitive task load (CTL) that was introduced in Section 2 of this paper. In layer 1 of the CLES tool, the tasks that the operator performs are observed in order to calculate an operator's cognitive load based on the CLT model. The following information is required for each task that the operator performs: the type of task, starting time, and finishing time.

As discussed in Section 2, the CTL model consists of three components: time occupied (TO), level of information processing (LIP), and task set switching (TSS). In layer 2 of the CLES tool, a value is calculated for each of these components for a given time frame, based on the observables represented in layer 1. TO is calculated by taking the percentage of time the operator was performing tasks in that time frame. Layer 2 contains domain knowledge representing the level of information processing for all tasks in that domain. LIP is calculated by taking the average level of information processing of all tasks performed in the time frame. TSS is the total number times the operator switched tasks in the time frame. Subsequently, these three values are combined to determine the cognitive load of the operator in that time frame (see [15] for the set of formula).

The cognitive load of an operator is continuously determined in layer 2. This value, changing over time, forms the input for layer 3 of the CLES tool. In this layer, sudden changes in cognitive load are detected (the next version will also detect risks for vigilance and cognitive lock-up). Furthermore, the combination of an operator's cognitive load and emotional state is being monitored and reasoned about.

3.3 Emotional State Modeling and Fuzzy Logic

Emotional state modeling is based on the Pleasure-Arousal-Dominance (PAD) model [16]. This model quantifies emotional state according to three dimensions: pleasure, arousal and dominance. An option is to leave out the dimension of dominance, resulting in the Valence-Arousal (VA) model which still produces a useful classification of emotional states

In layer 1 of the CLES tool, information is collected to determine an operator's emotional state. There are different ways to assess someone's valence and arousal. Valence can, for example, be assessed by recording and classifying someone's facial expressions, or by using sensors that measure facial muscle activity. Arousal is usually assessed by measuring someone's hart rate or hart rate variability, and galvanic skin response.

In layer 2, fuzzy logic is used to determine an operator's emotional state based on the physiological measures observed in layer 1 (see also [17]). We chose to use fuzzy logic because it enables reasoning with multiple statements that are partially true, rather than statements that are merely truth or false. Partial truth of a statement is represented by a value that ranges between 0 and 1. In fuzzy logic, a collection of such values forms a fuzzy set, and in a fuzzy inference process, rules are applied to this set to produce a new fuzzy set. For example, fuzzy sets of heart rate and galvanic

skin response can be combined to create a new fuzzy set of arousal, and fuzzy sets of arousal and valence can be combined to a new fuzzy set of emotional state. An example of a fuzzy inference rule is: 'if galvanic skin response is high and heart rate is medium high then arousal is high'. Finally, when all fuzzy sets are combined, the final fuzzy set undergoes a process of defuzzification to produce a single, most likely outcome. In the CLES tool this final outcome represents the operator's emotional state.

Like cognitive load, an operator's emotional state is continuously determined in layer 2, and used for the reasoning process in layer 3 of the CLES tool. The outcome of this reasoning can be, for instance, a warning for the operator or his superior when the operator is heavily burdened.

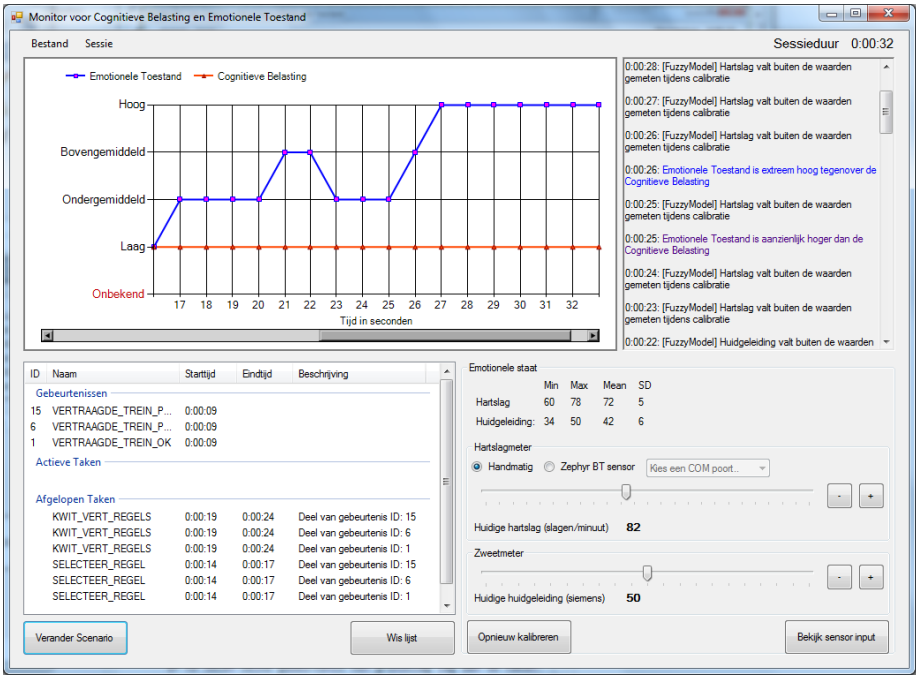


Fig. 1. Screenshot of the CLES monitor while running

4 The CLES Tool

Figure 1 shows a screenshot of the CLES tool while running. The display is divided into four components. The right and left parts in the lower half of the display give an overview of the operator's task performance and physical measures, respectively. These parts correspond to layer 1 in the CLES tool. The upper left part of the display shows the respective cognitive load (red line) and emotional state (blue line) of the operator. When the CLES tool is running, the graph is moving. This part corresponds to the output of the cognitive load and emotional state modeling in layer 2 of the

CLES tool. The upper right part of the displays shows messages that are the result of the reasoning component in layer 3 of the CLES tool. One of the messages is, for example, that emotional state is extremely high with regard to cognitive load.

In its current condition, the CLES tool perceives information about the heart rate and galvanic skin response of the operator. Input of heart rate information can be manually, through a slider, or automatically through a Zephyr BT sensor. Input of galvanic skin response is manually. The nature and number of input values can easily be adapted. Based on the physical measures, through fuzzy logic, the operator's level of arousal is determined. In the current CLES tool, thus, arousal is used to indicate emotional state.

Table 1. Example scenario of train breakdown

Scenario	**Explanation**
task(communication, 10, 100, DT)	Conversation with driver of defective train
task(communication, 110, 200, DT)	Conversation with traffic control officer
task (select rule, 200, 203, DT) task (find train, 205, 220, DT) task(automatic program off, 220, 223, DT) task(deselect rule, 223, 225, DT)	Cancel automatic control of defective train
task(recall signal, 230, 250, DT)	Recall signal at switch
task(select rule, 260, 263, DT) task(find train, 265, 285, DT) task(automatic program off, 285, 288, DT) task(manually process rule, 290, 350, DT) task(automatic program on, 350, 353, DT) task(deselect rule, 355, 358, DT)	Cancel automatic control
task(communicate, 380, 440, DT)	Conversation with mechanic
task(select rule, 440, 443, DT) task(find train, 445, 465, DT) task(manually process rule, 470, 530, DT) task(automatic program on, 530, 533, DT) task(deselect rule, 535, 538, DT)	Recover automatic control

Task performance information is generated by a simulator in the present state of the CLES monitor. The simulator simulates events such as a train that is delayed or a switch that does not work, and based on that, a list of tasks that the traffic control operator needs to perform. Each task is annotated with a beginning and an end time. An example scenario of a train that breaks down is provided below. The CLES tool has knowledge about the level of knowledge processing associated to these tasks, and uses this knowledge and the input from the simulator to determine the operator's cognitive load.

Table 1 presents an example scenario about a train that breaks down. Information is represented in the following way: task(id, t_{start}, t_{end}, event), where id refers to task id, t_{start} to the starting time of the task, t_{end} to the end time of the task, and event to the event for which the task was performed. In the scenario below, all tasks were performed in the context of event DT, a defective train. The left column shows the tasks of the operator, and the right column provides an explanation of the tasks.

5 Conclusions and Discussion

Based on prolonged research in different domains, we have identified, combined and formalized models of cognitive and affective load with transparent and easy-to-modify classification and assessment techniques. Fuzzy logic and its (membership) rules are generated to map a set of values to an emotional state, and to detect surprises of anomalies. The models and algorithms allow humans *to remain in the loop* of workload assessments and distributions, an important resilience requirement of human-automation teams. The current tool provides basic feedback, which is expected to improve human-automation team's awareness about the adequacy of the workload distributions and possibilities for improvement. By detecting unexpected changes (surprises and anomalies) and the corresponding cognition-emotion-performance dependencies, the CLES monitor can improve team's responsiveness to new situations, i.e., its resilience.

Current research focuses on the development of ePartners that are continuously informed by the CLES monitor. Based on this knowledge and other available information (e.g. from a user model), the ePartner will improve the feedback and provide advice at the individual and team level. The advice concerns, for example, the provision of insight in person's own functioning to improve his self-efficacy, and proposals for task (re)allocation to improve team performance [18].

Acknowledgement. This research was conducted within the RAILROAD project and is supported by ProRail and the Netherlands organization for scientific research (NWO) (under grant 438-12-306).

References

1. Ferreira, P.N., Balfe, N.: The contribution of automation to resilience in rail traffic control. In: Harris, D. (ed.) EPCE 2014. LNCS (LNAI), vol. 8532, pp. 469–480. Springer, Heidelberg (2014)
2. Hollnagel, E., Woods, D.D., Leveson, N. (eds.): Resilience engineering: concepts and percepts. Ashgate Publishing Limited, Hampshire (2006)
3. Neerincx, M.A., Grant, T.: Evolution of Electronic Partners: Human-Automation Operations and ePartners During Planetary Missions. Journal of Cosmology 12, 3825–3833 (2010)
4. Neerincx, M.A.: Situated Cognitive Engineering for Crew Support in Space. Personal and Ubiquitous Computing 15(5), 445–456 (2011)
5. Wickens, C.D., Hollands, J.G., Parasuraman, R., Banbury, S.: Engineering Psychology & Human Performance, 4th edn. Prentice-Hall (2012)

6. Picard, R.: Affective computing. MIT Press, Cambridge (1997)
7. Neerincx, M.A., Lindenberg, J.: Situated cognitive engineering for complex task environments. In: Schraagen, J.M.C., Militello, L., Ormerod, T., Lipshitz, R. (eds.) Naturalistic Decision Making and Macrocognition, pp. 373–390. Ashgate Publishing Limited, Aldershot (2008)
8. Rasmussen, J.: Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering. Elsevier, Amsterdam (1996)
9. Neerincx, M.A., Griffioen, E.: Cognitive task analysis: harmonizing tasks to human capacities. Ergonomics 39(4), 543–561 (1996)
10. Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A.M., Koch, I.: Control and interference in task switching—A review. Psychological Bulletin 136(5), 849–874 (2010)
11. Neerincx, M.A.: Cognitive task load design: model, methods and examples. In: Hollnagel, E. (ed.) Handbook of Cognitive Task Design, ch. 13, pp. 283–305. Lawrence Erlbaum Associates, Mahwah (2003)
12. Grootjen, M., Neerincx, M.A., Veltman, J.A.: Cognitive task load in naval ship control centres: from identification to prediction. Ergonomics 49, 1238–1264 (2006)
13. Neerincx, M.A., Kennedie, S., Grootjen, F., Grootjen, M.: Modelling Cognitive Task Load and Emotion for Adaptive Maritime Interfaces. In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience. Proceedings of the 5th International Conference of the Augmented Cognition. LNCS (LNAI), pp. 260–269. Springer, Heidelberg (2009)
14. Zadeh, L.A.: Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers, vol. 6. World Scientific (1996)
15. Harbers, M., Aydogan, R., Jonker, C.M., Neerincx, M.A.: Sharing Information in Teams: Giving Up Privacy or Compromising on Team Performance? In: Proceedings AAMAS 2014, Paris, France, May 5-9 (2014)
16. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology 14(4), 261–292 (1996)
17. Mandryk, R.L., Atkins, M.S.: A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human-Computer Studies 65(4), 329–347 (2007)
18. De Greef, T.E., Arciszewski, H.F.R., Neerincx, M.A.: Adaptive Automation based on an Object-Oriented Task Model: Implementation and Evaluation in a Realistic C2 Environment. Journal of Cognitive Engineering and Decision Making 4, 152–173 (2010)

Multitasking and Mentalizing Machines: How the Workload Can Have Influence on the System Comprehension

Oronzo Parlangeli[1], Maria Cristina Caratozzolo[2], and Stefano Guidi[2]

[1] Department of Social, Political and Cognitive Sciences, University of Siena
parlangeli@unisi.it
[2] BSD Design
{cristina.caratozzolo,stefano.guidi}@bsdesign.eu

Abstract. The laboratory study we are carrying out is aimed at discovering possible correlations between multitasking activity, workload and the attribution of mental states to technological systems. The scores of mental states attribution provided by subjects allotted to three different experimental conditions (one task, two concurrent tasks, three concurrent tasks) have been compared. Preliminary results show an increase in the tendency to attribute mental states as the operational workload increases.

Keywords : multitasking, attribution of mental states, workload.

1 Introduction

Research in cognitive ergonomics has often taken for granted that human-computer interactions are based on the elaboration of some interpretative models. That is, expectations about the way a system works come from the mental models progressively elaborated by the user through the consideration of the computer behavior. These models are traditionally considered as deterministic models of processes, sometimes they are incomplete or not coincident with the actual functioning of the system, but quite often they are sufficiently adequate to give course to a productive interaction.

More specifically, the user is believed to elaborate mental models that would be used as a means to make predictions, to produce explanations, and to provide diagnosis about the behavior of the system (Allen, 1997).

As a consequence of that, the communication between man and machine is in fact set on the basis of the understanding that the user is able to build about the system, and will therefore be much more efficient, as the latter will be more accurate.

The common experience is that human beings generally interact with inanimate systems making use of an implicit knowledge of proper physical laws. Though, in some cases that have been extensively investigated, the interaction with many mechanical systems, and particularly with information and communication technologies, seems to be based on other interpretative rules (Molina et al., 2004). It

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 50–58, 2014.

is well known that people tend to consider as human agents those systems that move and/or show some changes in even simple characteristics, such as shape, color, and size (Dittrich 1994, Morewedge 2007). This bias, that is the liability to consider human-made systems as if they were human beings, seems to depend on conceptualizing these systems as if they were gifted with some self-generated and self-controlled cognitive ability (Epley et al. 2007, Kelemen, Carey 2007, Terada 2007).

It seems quite clear that this phenomenon involves one of the human tendencies that is probably amongst the most surprising and advantageous from an evolutionary point of view, namely the bias that brings us to attribute mental states, to elaborate a theory of mind (Premack and Woodruff, 1978; Dennett, 1987), to and for nearly all the entities with which we engage in some kind of interaction.

In the last years, the human tendency to anthropomorphize - in this context it could be said "mentalize" - nearly everything, has been gaining increasing attention. For what concerns the explanation of how a theory of mind is developed by human beings since their birth, it is possible to identify two opposite hypoteses, the one seeing this tendency as innate (Baron-Cohen, 1995; Premack, 1990; Perrett and Emery, 1994) and the one framing it as a competence that is structured mainly through actual experiences (Meltzoff, 1995; Tomasello 1999).

Some recent studies (Steinbeis and Koelsch, 2009) in the field of neuropsychology have shown that when people believe that they are interacting with an artifact, that is with a human product, it is possible to record a cortical activity that is in the same cortical network (anterior medial frontal cortex – superior temporal sulcus – and temporal poles) that is usually activated during processes of mental states attribution

In the field of human-computer interaction, however, this issue has never received much attention. This in spite of the fact that understanding the way in which users elaborate a theory of mind for what concerns computer behaviors could be clearly very useful to design and implement more user-friendly technological systems.

Some studies have been conducted in order to investigate which determinants can induce the adoption of a theory of mind in relation to the behavior of some robots – technological systems that often, even in their appearance, can closely resemble human beings. In these cases, the studies have generally supported the hypothesis that considers human beings more prone to the attribution of mental states if the interactive systems exhibit actions that are reactive to user behavior, and if their affordances can be more easily detected.

Overall, however, there is still a surprising lack of knowledge about the phenomenon of mental states attribution to artificial complex systems. So far, for instance, we do not know whether the attribution of mental states is an all-or-nothing process, or whether different mental states, such as intentionality and awareness, are seen linked together in the process of attribution of a mental entity. It is also actually unclear if some contextual variables, that are neither inherent to the user nor to the system, may affect the occurrence of such a phenomenon.

2 Multitasking

Multitasking can be described as the behavior that allow people to cope with more than one task at a time. Research has recently provided evidence that during the last decades, likely due to the increased availability of technological systems, multitasking has become a very common behavior, and it is relatively more common among the younger generation (Roberts et al. 1999; Foehr 2006).

Reasons for engaging multitasking activities have not only been related to the growth in number of the technological systems. Other theoretical perspectives have focused on the psychological determinants for multitasking. Different authors (Albarran et al. 2006; Pornsakulvanich, et al. 2008; Zhang and Zhang, 2012) have referred to the theory of uses and gratification as an explanatory hypothesis for multitasking while interacting with ICTs. In this perspective, gratifications are considered as one of the most relevant factors in shaping human-computer interaction. More recently Sanbonmatsu et al. (2013) have suggested that those people who are more prone to engage in multitasking activities are also less able to block out distractions and to dedicate all their attentional resource to a single task.

Then, the willing of an individual to undertake multitasking activities with technological systems probably depends either on contextual factors, such as the availability of technologies, and on psychological factors, such as the control of his cognitive resources.

In a reference to the use of cognitive resources, an obvious effect – often particularly emphasized by popular science – is highlighted: multitasking activities can erode cognitive resources in a consistent manner. It follows that in multitasking activities the performance of each individual task can degrade until the occurrence of the condition in which different tasks, contending the same resources, cannot be executed properly.

Is now a widely accepted hypothesis that for the execution of various tasks the same systems and the same cognitive processes can be committed.

Just think of the enormous deal of research that, basing on the evidence of interference in the performance of dual tasks, led to the development of theories such as those concerning the existence of different subordinate systems referring to working memory (Baddeley and Hitch, 1974). Well-known findings on this field tell us that tasks that must be performed at the same time can be especially difficult if, for example, they are similar to each other (Treisman and Davies, 1973) and if the subjects are not experienced, and therefore the necessary tasks for their planning and execution have not been automated (Everett, 2011).

According to these considerations, may be therefore informative to know whether subjects that are more likely to engage in multitasking activities with technological systems are more or less inclined to attribute mental states to technologies; namely, to clarify if and how an augmentation of the workload (not only at a cognitive level) has an effect on the processes that lead to the discrimination of mental agents from those that are not.

As a matter of fact, this could be illuminating, for example, on the role of the attentional processes in the process of anthropomorphization or in the acquisition of an intentional stance. It may also suggest some hypotheses on the development of the discerning capacity underlying the mental processes that discriminate intentional agents from those that are not.

Finally, it could provide indications that the younger generation, given the large amount of time they spend interacting simultaneously with multiple technologies, might live in a world that is populated by systems that are perceived as more or less intentional than their parents do.

Trying to answer these questions, a laboratory study is being carried out; it involved 60 subjects. In the following we will give account of this study, trying to report the most important preliminary results.

3 The Study

The laboratory study we are carrying out is aimed at discovering possible correlations between the accomplishment of a multitasking activity, the increase of workload and the attribution of mental states to technological systems.

In order to pursue this aim, the experimental session is structured in three different phases: an initial questionnaire, the actual task, and finally another questionnaire, for a total duration of approximately forty minutes.

To begin, the subjects are invited to provide some general socio-demographic information and some indications on their use of media tools and applications.

Then, a tool developed by Ophir et al (2009) is used to deduct if and how much the subjects use different media simultaneously. The media taken into consideration are: Social media, TV, computer videos, music, video games, telephone, instant messaging, text messaging, e-mail, web, other computer applications.

The subjects must complete a matrix in which each of the above-mentioned media is considered as the primary mean: They have to report how often they use simultaneously (as a secondary mean) each one of the other media (see Figure 1).

Thanks to the information provided in this matrix, it is possible to derive the Media Multitasking Index (MMI), which defines at what level the subject is – or is not – a multitasker.

After the pre-test questionnaire, the subjects are asked to perform the proper laboratory test; they are randomly allotted in three groups, and requested to perform tasks of increasing complexity. The first group faces a, quite simple, single task, while the second and third group have to perform a multitasking activity – two and three tasks at the same time, respectively.

Scrivi usando dei numeri per indicare le parole: **1 – mai** **2 – raramente** **3 - a volte** **4- spesso**

attività principale / *attività secondaria*	social network	Guardi la TV	Guardi dei video al computer	Ascolti musica	videogiochi	Parli al telefono	instant messaging	sms	e-mail	web	Altri programmi
Interagendo sui social network											
Guardando la TV											
Guardando dei video al computer											
Ascoltando musica											
Facendo dei videogiochi											
Parlando al telefono											
Usando programmi di instant messaging											
Scrivendo o leggendo sms											
Scrivendo o leggendo e-mail											
Navigando sul web											
Usando qualsiasi altro programma al computer											

Fig. 1. The Italian version of the Ophir et al (2009) tool, used to define the MMI index

To be specific: The subjects of the first group only see the squares on the screen (refer to Figure 2), which change their colour every 2 seconds; they only have to press a key if at least 3 out of 4 of the rectangles are the same colour.

Fig. 2. The stimuli of different experimental conditions

The subjects in Condition 2 can also see the lines of a poem of the XIX century (two hendecasyllables that remain on the screen for 4 seconds) and, in addition to performing the task of Condition 1, they must also read those verses aloud.

The subjects in Condition 3 perform the previous two tasks and, moreover, they have to click with the mouse when they read a verse with at least one comma.

In all three conditions, the duration of the whole task is about 7 minutes and 30 seconds.

The third phase of the experimental session consists of two steps. First, the subjects are required to complete a questionnaire concerning the attribution of mental states: They are asked to report if, in the course of the interaction, they happened to think that the coloured rectangles/the application had: Awareness, their own strategy, intentions, a mind, capability for attention, recollections, etc.

Finally, in order to verify that an increase of the number of simultaneous tasks also increases the workload, the subjects must also complete the NASA Task Load Index (NASA-TLX): It is a subjective workload assessment tool developed by the Human Performance Group at NASA's Ames Research Center. It allows users to perform subjective workload assessments on operators working with various human-machine systems, through a multi-dimensional rating procedure that derives an overall workload score based on a weighted average of ratings on six subscales.

These subscales include: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort and Frustration.

4 Results

The preliminary results of this study are showing a clear relationship between performance of multitasking and taking an intentional stance. We compared the subjects' believes about the system having mental states across the different experimental conditions, using both MANOVA and non-parametric tests. Both kinds of tests showed significant differences between the conditions for 5 mental states attributions: Awareness ($p<0.05$), Intentions ($p<0.05$), Attention ($p<0.01$), Memories ($p<0.05$), Mind ($p<0.05$). As a matter of fact, the subjects reported that they attributed mental states to the rectangles on the screen with greater ease and frequency, as the number of tasks they had to perform increased (Figure 3). Pairwise comparisons, however, for most mental states showed significant differences only between the control condition (no multitasking) and the double multitasking condition, and no differences between the two different multitasking conditions, although this could be due to a lack of statistical power in this early stage of the study. As a matter of fact, the analysis of the mean NATA-TLX scores across the conditions, confirmed that indeed the perceived workload significantly increased with the number of tasks to be performed ($p<0.0001$), but also showed significant differences in all the pairwise comparisons.

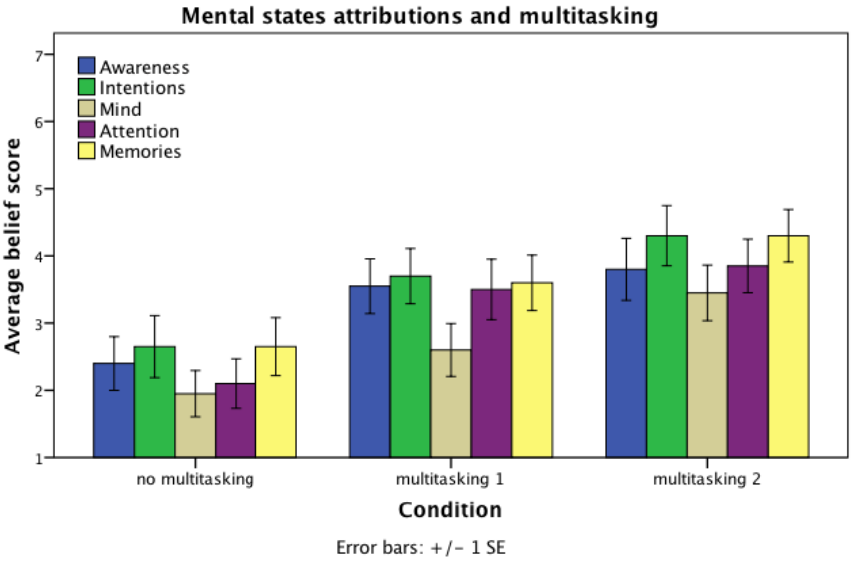


Fig. 3. Bar plots represents the average belief scores for five different mental states, as function of condition (i.e. presence and degree of multitasking). Error bars are standard errors.

This willingness to attribute mental states, finally, does not seem to be correlated with the MMI index, namely the individual tendency to undertake multitasking. It therefore seems possible to conclude that the more a person is induced to operate in multi-tasking activities, the more will be brought to believe that the technologies with which they are dealing with mental states.

5 Discussion

In his seminal work Dennett (1987) has put forward that the choice of which particular stance must be adopted depends on factors as the level of accuracy that the task at hand requires to be properly performed, and by how successful that stance has resulted in similar circumstances when formerly applied.

Several studies, however, have shown that both contextual and personal factors can affect and change the tendency to attribute mental states to the artifacts and technologies surrounding us.

In a previous study (Parlangeli et al. 2013) the authors noted a clear relationship between the multitasking and the attribution of mental states to technological systems: the more a subject declares him/herself as a multitasker, the more he/she reports of having experienced circumstances in which he/she thought that his/her computer had mental states.

An open question remained, however, about whether this is due to the fact that individuals who are more likely to undertake multitasking activities are also more willing to assume an intentional stance with regard to the technological systems, or

that it is the multitasking activity itself, requiring a considerable commitment of the cognitive resources, brings to an easier attribution of mental states to technologies contextually.

Consequently, the present study opens new perspectives of interpretation of this phenomenon. On the one hand, it follows that the subjects that are frequently pursuing a multitasking are more inclined to attribute mental states to the technologies they use. On the other hand, this data are further defined by the reference to the mental workload required to perform these tasks, and probably to a considerable use of attentional resources: Individuals who are brought to operate in multitasking mode are more inclined to attribute mental states to technologies.

This allows us to suppose that the ability to discriminate mental from non-mental agents can be partially weaken by our attentional resources being maximally involved in the execution of multiple tasks. As if to say that, when we are particularly committed by a cognitive point of view, perhaps we fail to assume a correct, but costly, rational attitude.

It also seems possible that, in maximum operational commitment circumstances, less evolved cognitive procedural rules could emerge, rules that are not able to lead to a fine discrimination between mental and non-mental entities.

References

1. Albarran, A.B., Anderson, T., Bejar, L.G., Bussart, A.L., Dagget, E., Gibson, S., et al.: What Happened to our Audience? Radio and New Technology uses and Gratifications among Young Audult Users. Journal of Radio Studies 14(2), 2–11 (2006)
2. Allen, R.B.: Mental models and user models. In: Helander, M., Landauer, T.K., Prabhu, P. (eds.) Handbook of Human-Computer Interaction. Elsevier Science (1997)
3. Baddeley, A.D., Hitch, G.: Working Memory. In: Bower, G.A. (ed.) Recent Advances in Learning and Motivaation. Academic Press, New York (1974)
4. Baron-Cohen, S.: (1) Mindblindness: An Essay on Autism and Theory of Mind. The MIT Press, Cambridge (1995)
5. Baron-Cohen, S.: (2) The eye direction detector (edd) and the shared attention mechanism (sam): Two cases for evolutionary psychology. In: Moore, C., Dunham, P.J. (eds.) Joint Attention: Its Origins and Role in Development, vol. 3, pp. 41–59. Lawrence Erlbaum Associates (1995)
6. Dennett, D.C.: The Intentional Stance. Bradford Books/MIT Press, Cambridge, Mass. (1987)
7. Dittrich, W.H., Lea, S.E.G.: Visual perception of intentional motion. Perception 23(3), 253–268 (1994)
8. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: A three-factor theory of anthropomorphism. Psychological Review 114(4), 864–886 (2007)
9. Everett, J.: Implications from Dual Tasking Research: can we really do two things at once? Psychtalk 70 (2011)
10. Foher, U.G.: Media Multitasking among American Youth: Prevalence, Predictors, and Pairings. Kaiser Family Foundation, Menlo Park (2006)
11. Kelemen, D., Carey, S.: The essence of artifacts: Developing the design stance. In: Lawrence, S., Margolis, E. (eds.) Creations of the Mind: Artifacts and their Representation. Oxford University Press, Oxford (2007)

12. Meltzoff, A.N.: Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. Developmental Psychology 31(5), 838–850 (1995)
13. Molina, M., Van de Walle, G.A., Condry, K., Spelke, E.S.: The animate–inanimate distinction in infancy: Developing sensitivity to constraints on human actions. Journal of Cognition and Development 5, 399–426 (2004)
14. Morewedge, C.K., Preston, J., Wegner, D.M.: Timescale bias in the attribution of mind. Journal of Personality and Social Psychology 93(1), 1–11 (2007)
15. Ophir, E., Clifford, N., Wagner, A.D.: Cognitive control in media multitaskers. Proceedings of the National Academy of Sciences of the United States of America (September 5, 2009)
16. Parlangeli, O., Guidi, S., Caratozzolo, M.C.: The attribution of mental states to technological systems. Paper Presented at IV Joint Workshop Rutgers-Siena on Cognitive Sciences, RUCCS, Rutgers University (May 21, 2013)
17. Pornsakulvanich, V., Haridakis, P., Rubin, A.M.: The influence of dispositions and Internet motivation on online communication satisfaction and relationship closeness. Computers in Human Behavior 24(6), 2292–2310 (2008)
18. Perrett, D.I., Emery, N.J.: Understanding the intentions of others from visual signals: Neurophysiological evidence. Current Psychology of Cognition 13, 683–694 (1994)
19. Premack, D.: The infant's theory of self-propelled objects. Cognition 36(1), 1–16 (1990)
20. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? The Behavioral and Brain Sciences 4, 515–526 (1978)
21. Roberts, D.F., Foher, U.G., Rideout, V.J., Brodie, M.A.: Kids and Media at the New Millennium. Kaiser Family Foundation, Menlo Park (1999)
22. Terada, K., Shamoto, Y., Mei, H., Ito, A.: Reactive Movements of non-humanoid robots cause intention attribution in humans. In: Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA (2007)
23. Sanbonmatsu, D.M., Strayer, D.L., Medeiros-Ward, N., Watson, J.M.: Who Multi-Tasks and Why? Multi-Tasking Ability, Perceived Multi-Tasking Ability, Impulsivity and Sensation Seeking. PLoS ONE 8(1), e54402 (2013), doi:10.1371/journal.pone.0054402
24. Steinbeis, N., Koelsch, S.: Understanding the Intentions behind man-made products elicits Neural Activity in area dedicated to mental state Attribution. Cerebral Cortex 19(3), 619–623 (2009)
25. Tomasello, M.: The cultural origins of human cognition. Harvard University Press (1999)
26. Treisman, A., Davies, A.: Dividing attention to ear and eye. In: Kornblulm, S. (ed.) Attention and Performance IV, pp. 101–117. Academic Press, New York (1973)
27. Zhang, W., Zhang, L.: Explicating Multitasking with Computers: Gratifications and Situations. Computers in Human Behavior 28, 1883–1891 (2012)

Neuronal Mental Workload Registration during Execution of Cognitive Tasks

Thea Radüntz

Federal Institute for Occupational Safety and Health,
Unit. 3.4 'Mental Health and Cognitive Capacity'
Nöldnerstr. 40/42, D-10000 Berlin, Germany
raduentz.thea@baua.bund.de

Abstract. Neuronal workload measurement is a key-technology for optimizing work conditions in human-machine systems. Specific aims are the identification of neurophysiological parameters indicative for workload and their validation by systematic variation of external load conditions.

The battery consists of tasks with diverse complexity and difficulty. The sample consists of 34 people and shows high variability in respect to the cognitive capacity and hence to the experienced mental workload. The electroencephalogram (EEG) as well as further workload relevant bio signal data and the NASA-TLX as a subjective questionnaire method are registered.

Results from the NASA-TLX questionnaire reveal the predominant role of the mental dimension at the implemented task battery. Furthermore, the NASA-TLX indicates the existence of diverse levels of difficulty with several tasks per level. Analysis of EEG spectra demonstrates an increase of frontal theta band power and a decrease of alpha band power with increasing task difficulty level.

Keywords: mental workload, electroencephalogram (EEG), signal processing, pattern recognition.

1 Introduction

The development of advanced information and communication technology as well as of highly interactive work environments and work assistance systems is unstoppable. Although the main goal of this development is to simplify the work, employees complain about high mental workload and stress. Problems arise from information overload, frequent work interruptions or from a multitude of irrelevant information [9], [10], [13].

On the other hand work associated with automation and supervisory control can be linked to repetitive and monotonous tasks that may be accompanied by complacency, fatigue, reduced vigilance [14], [15], [3], [2], [12] and increased error rates, hence a safety risk for further persons [17].

An objective method for mental workload registration is absolutely essential particularly with regard to the long-term negative consequences of inappropriate

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 59–69, 2014.

workload on the individual's health as a serious problem of modern society. Appropriate work efficiency is only possible in an optimal workload range that can most efficiently be measured where information processing takes place, i.e. in the brain. Neuronal workload measurement is hence a key-technology for optimizing work conditions in human-machine systems.

Hence, the overall goal is continuous, online registration and monitoring of mental workload on neuronal basis. The theoretical background is given by the variability of the EEG frequency bands according to attention, fatigue and mental workload. Specific aims are the identification of neurophysiological parameters indicative for workload and their validation by systematic variation of external load conditions.

2 Methods

The tests took place in the shielded lab of the Federal Institute for Occupational Safety and Health in Berlin.

The electroencephalogram (EEG), as a direct signal of bioelectrical brain activity, as well as further workload relevant bio signal data (e.g. heart rate, blood pressure) and the NASA-TLX as a subjective questionnaire method are registered. Hence, subjective and objective methods can be combined and contribute to the validation of the mental workload registration.

2.1 Procedure

The experiment was fully carried out with each subject in a single day. It consisted of two parts: a training phase and the main experiment. During the training phase subjects were familiarized with the cognitive tasks. The cognitive tasks were the same as these of the main experiment but shorter in time. They were repeated until the subject reached an accuracy index of at least 80%. The training phase should create similar individual starting conditions in respect to the performance, so that we can investigate the workload's effect independent from learning effects.

The main experiment started after a short break subsequent to the training phase. The tasks were presented in the same counterbalanced order as presented during the training phase. They took place in the shielded lab of the Federal Institute for Occupational Safety and Health and were controlled remotely through a remote desktop connection, an intercommunication system and a video monitoring system.

2.2 Subjects

The sample consists of 57 people in paid work and shows high variability in respect to the cognitive capacity and hence to the experienced mental workload. At the time of writing only the data of the first 34 people has been analyzed. Table 1 describes the sample set used.

Table 1. Sample set

Age	Male	Female	Total
30 - 39	4	5	9
40 - 49	7	8	15
50 - 59	1	5	6
60 - 69	3	1	4
Total	15	19	34

2.3 Tasks

The simulation of miscellaneous cognitive task requirements is realized through the implementation of a task battery in the E-Prime application suite. The battery consists of tasks with diverse complexity and difficulty inducing different levels of mental workload. The implemented tasks are listed in Table 2.

Table 2. Task battery

Task	0nb	2nb	Sternberg	Seriell sternberg	Stroop
duration [m]	5	5	10	10	5
Task	Switch PAR	Switch NUM	Switch XXX	AOSPAN	
Duration [m]	5	5	10	20	

In this paper we concentrate on the analysis and evaluation of three tasks: 0-back as the easiest one, 2-back as a working memory task with moderate workload, and aospan as a demanding dual task (see Figures 1, 2, 3). The latter was the self adapted and translated version of the AOSPAN task developed by [18].

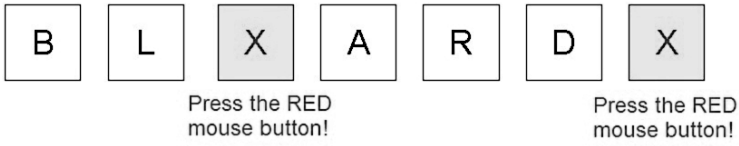


Fig. 1. 0-back task: Press the mouse button if the presented letter is 'X'

2.4 Subjective Ratings

Subjective workload was captured with a computerized version of the NASA-TLX [6]. After each task during the training phase, subjects were asked to rate the workload sources in 15 pairwise comparisons of NASA-TLX's six workload dimensions: mental demand, physical demand, temporal demand, performance,

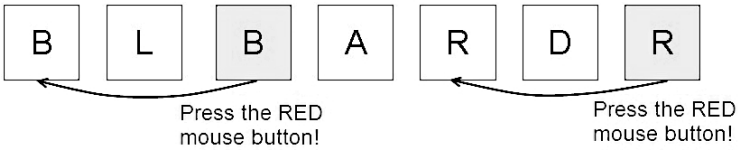


Fig. 2. 2-back working memory task: Press the mouse button if the presented letter is the same as the next to last letter seen.

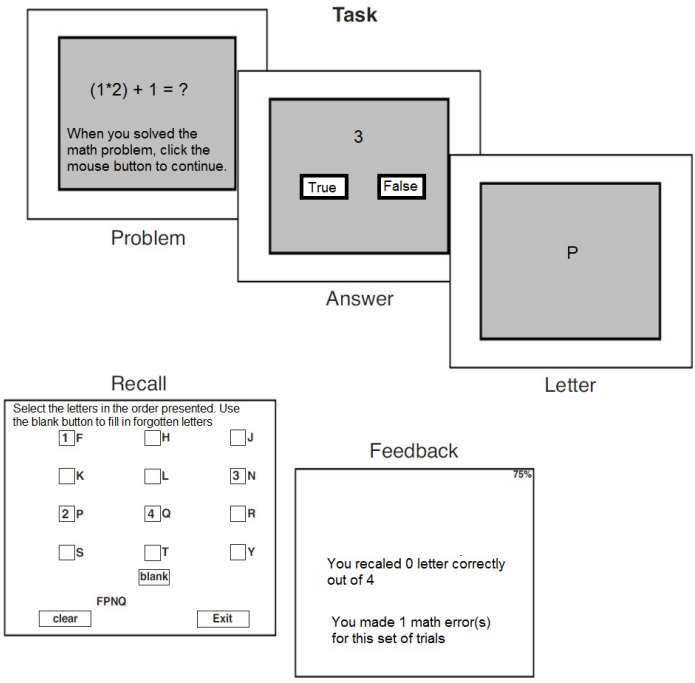


Fig. 3. AOSPAN dual task: memorize letters in the order presented while simultaneously solving math problems. Trials consist of 3 sets of each set-size, with the set-sizes ranging from 3-7.

effort, frustration. This required the subject to choose which dimension is more relevant to workload in the specific task. Hence, we gained an individual weighting of these subscales based on their perceived importance.

After each task during the main experiment, subjects were asked again to rate the task within a 100-points range with 5-point steps. They indicated their rating by clicking on a 5-point step box with an optical mouse.

2.5 Physiological Measures

The electroencephalogram (EEG) as well as the blood pressure (BP), the heart rate (HR) and the inter-beat interval (IBI) were digitally recorded during the main task only. Digital signal processing and calculation of mean values were done with MATLAB.

EEG. The EEG was captured by 25 electrodes placed at positions according to the 10-20-system and recorded with reference to Cz and at a sample rate of 500 Hz. For signal recording we used an amplifier from BrainProducts GmbH and their BrainRecorder software.

For workload calculations we implemented a modular MATLAB-Toolbox. Figure 4 describes the signal processing pipeline.

The pre-processing module reads the recorded EEG signal. The signal is multiplied with a Hamming window function and filtered with a bandpass filter (order 100) between 0.5 and 40 Hz. Subsequently, independent component analysis (ICA) is applied to the signal and the calculated independent components are visually inspected and classified as either an artifact or signal component. The signal components are projected back onto the scalp channels and the now artifact-corrected EEG signal is passed over to the next module. There it is cut into segments of 10 seconds length, overlapping by 5 seconds. The segments are then transformed to frequency domain using Fast Fourier Transformation (FFT) and workload relevant frequency bands are computed (θ: 4-8 Hz, α: 8-12 Hz).

The combination of the θ- and α-band provides the basis for the indexing, training and classification of mental workload according to Lei's Logistic Function Model [11]:

$$W = \frac{1}{1 + e^{-(b0+b1*\theta+b2*\alpha)}} \quad . \tag{1}$$

The model allows the identification of individual workload.

Cardiovascular Parameters. Blood pressure was recorded continuously by the FMS Finometer Pro device. A finger cuff was placed around subject's finger and systolic and diastolic blood pressure as well as the heart rate and the inter-beat interval were detected automatically. The recorded data was processed in the time domain.

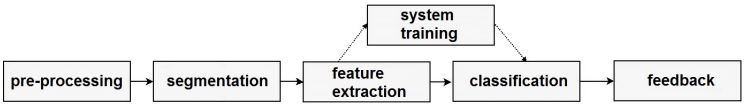


Fig. 4. Signal processing pipeline

2.6 Performance

Mean reaction time computation was meaningful only for 0-back and 2-back, where subjects were asked to respond quickly after the presented stimuli. During AOSPAN's recall slide there was no time pressure and the reaction time depended also on the set size presented and on mouse movements. In addition, subjects were allowed to correct themselves several times. The time alloted for solving the math tasks was computed individually during the training phase by calculating the mean time of the responses. In the main experiment, when this individually calculated time was exceeded, the next slide was shown and a math error registered for the current operation.

For all three tasks the individual percentage of false responses was calculated. For AOSPAN, false responses include the number of sets in which the letters are not recalled in correct serial order, math errors and the above mentioned math speed errors.

2.7 Analysis

Physiological measures were obtained and calculated for the following three tasks: 0-back, 2-back and AOSPAN. Six ANOVAs were carried out utilizing repeated measures design, one within-subject factor (θ, α, systolic BP, IBI, percentage of errors, NASA-TLX) and 3 levels (the tasks), where the differences between the tasks were examined and tested with a post-hoc test (Bonferroni). A paired-samples t-test was computed for the 0-back and 2-back reaction times.

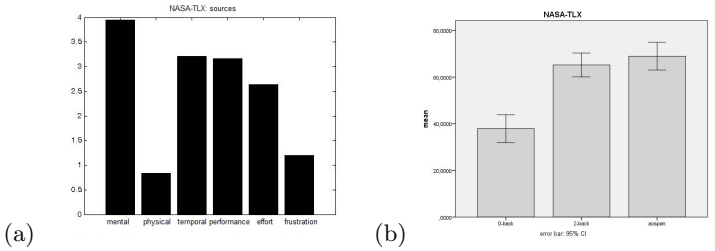
3 Initial Results

First results computed over 34 subjects and three tasks will be presented in the following section. They comprise the obtained subjective ratings and task performance as well as the workload relevant θ- and α-band from the EEG and the systolic BP and IBI.

3.1 Subjective Ratings

Results of the subjective ratings are presented in Figure 5. Figure 5(a) indicates the predominant role of mental demands at the implemented task battery. Hence, the induced workload originates from information processing and should

be reflected in the EEG. Figure 5(b) shows the average workload index for the selected tasks 0-back, 2-back and AOSPAN as representatives of a low, moderate and high workload tasks. Workload means changed significantly during the experiment (Greenhouse-Geisser $F(1.9;65.5) = 57.1$, $p<0.001$). Post-hoc analysis showed that the mean workload was significantly lower during the 0-back task than in the other two tasks.



(a) (b)

Fig. 5. (a) NASA-TLX: source rating over all tasks and 34 subjects. (b) NASA-TLX: workload index computed for 0-back, 2-back and AOSPAN over 34 subjects.

3.2 Physiological Measures

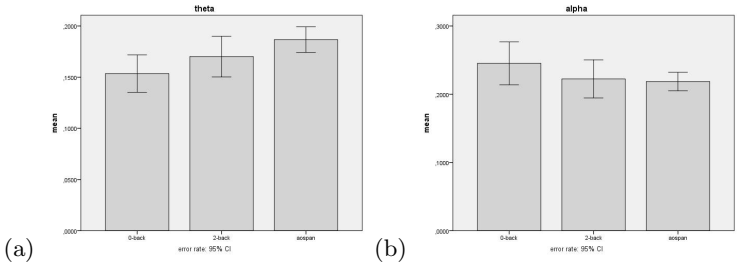
EEG. Analysis of the EEG spectra at the Fz and Pz electrode demonstrates an increase of the θ-band power and a decrease of the α-band power with increasing task difficulty level. This fact assures successful system training and an individual parametrization in the context of Equation 1.

Results obtained from the Fz electrode are presented in Figure 6. They are consistent with previous observations of several other authors (e.g. [16], [5], [4]). θ- and α-band means changed during the experiment. The θ-band changed significantly whereas the α-band revealed no significant changes (Greenhouse-Geisser $F(1.6;46.2) = 8$, $p<0.01$; Greenhouse-Geisser $F(1.9;63) = 3.6$, $p=0.04$). Post-hoc analysis of the θ-band showed that the means were significantly larger during the AOSPAN-task than in the 0-back task.

Cardiovascular Parameters. Both systolic BP and IBI differed between the three tasks significantly (Greenhouse-Geisser $F(1.3;45) = 15$, $p<0.001$; Greenhouse-Geisser $F(1.4;46.5) = 6.5$, $p<0.01$). IBI in the 0-back task were, according to post-hoc analysis, lower than in the 2-back and AOSPAN task. Systolic BP means were significantly larger during the AOSPAN-task than in 0-back and 2-back tasks. Results of systolic BP and IBI are presented in Figure 7(a) and Figure 7(b).

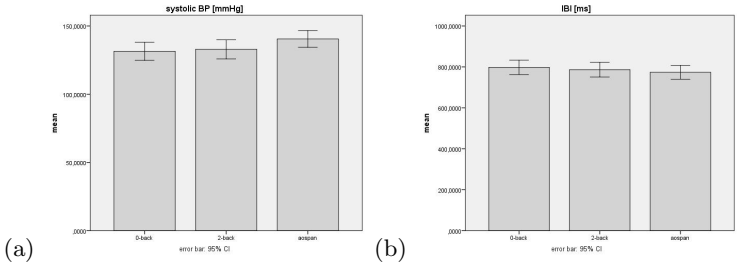
3.3 Performance

Mean reaction times (RT) and error rates are presented in Figure 8. Figure 8(a) demonstrates differences in RT between the 0-back and the 2-back task (t-test



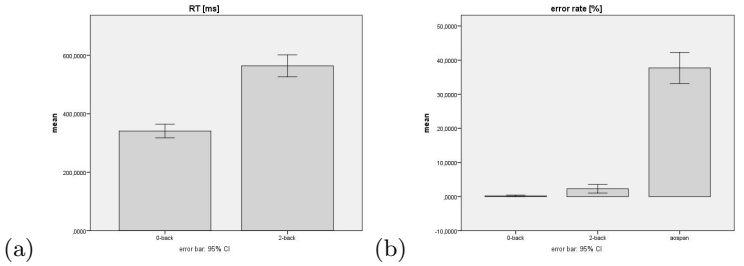
(a) (b)

Fig. 6. EEG - Fz: θ-band (a) and α-band (b) computed for 0-back, 2-back and AOSPAN over 34 subjects



(a) (b)

Fig. 7. Systolic BP (a) and IBI (b) computed for 0-back, 2-back and AOSPAN over 34 subjects



(a) (b)

Fig. 8. (a) Reaction time computed for 0-back and 2-back over 34 subjects. (b) Percentage of false responses computed for 0-back, 2-back and AOSPAN over 34 subjects.

$p<0.01$). Figure 8(b) shows the average error rate for the selected tasks 0-back, 2-back and AOSPAN. Error rate means changed significantly during the experiment (Greenhouse-Geisser $F(1.1;39.2) = 271.8$, $p<0.001$). Post-hoc analysis revealed significant changes of the mean percentage of false responses between all three tasks.

4 Discussion

The central issue addressed by this paper is the registration of workload by means of neurophysiological parameters. Therefore a task battery of miscellaneous cognitive task requirements was implemented, including various complexity and difficulty, hence inducing different levels of mental workload. In this paper we concentrated on the 0-back, 2-back and AOSPAN tasks as representatives for an easy, moderate and difficult task. Subjective ratings derived from the NASA-TLX questionnaire demonstrate significant workload differences only between the easy 0-back task and the other two.

The RT between 0-back and 2-back task increased significantly and also the error rate between the three tasks grew significantly. In contrast to the subjective ratings that could distinguish only between two levels, the performance score indicates a difference between all task difficulty levels. Furthermore, the computed error rate for the AOSPAN task is remarkably high.

The IBI, similar to the subjective ratings, shows significant differences only between the 0-back task and the other two, while the systolic BP means were significantly higher between the AOSPAN and the other two tasks and coincide with AOSPAN's extremely high error rate.

The EEG as a direct signal of brain activity and the frequently observed variability of the θ- and α-band according to attention, fatigue and mental workload, constitute the theoretical background for the implementation of an objective method for neuronal mental state monitoring. Initial results of our EEG signal processing demonstrate that the θ-band increases with advanced task difficulty. It can significantly distinguish between an easy task like 0-back and a difficult task like AOSPAN. The α-band shows the tendency to decrease with task difficulty increase but did not obtain significance. These observations are consistent with several other studies. There the θ-band is a reliable parameter which is enhanced with increasing difficulty, whereas the α-band seems to be less reliable with respect to the decrease which is normally expected. In some studies this behavior is linked to different forms of attention (internal vs. external) and other task requirements [8], [7]. However, the expected tendencies for an increase of the θ- and a decrease of the α-band are given in our study.

In addition and with respect to the gained error rates, we have to ask to which extent subjects remained sufficiently motivated during the AOSPAN task. A second question would be whether the subjects continued to invest enough effort in problem-solving, even if they realized that the demands on them exceeded their own processing capacity. That could also be a reason why subjects' ratings between 2-back and AOSPAN task did not reach significance. The systolic BP shows a significant increase during the AOSPAN task, but this could be linked to emotional reactions like frustration. It is commonly known that BP is influenced by this [1]. Furthermore, we quickly checked the NASA-TLX frustration subscale and noted that AOSPAN received the largest mean value there. The analysis of the battery's further tasks could help solving these questions. The same applies to the processing of the further subjects and the detailed analysis of NASA-TLX subscales.

Based on such consolidated findings of neuronal brain states an optimal task sharing between human and machine with efficient cognitive processing for the operator could be defined. The benefits for older employees are maintenance of autonomy and working ability due to the moderate mental workload when working with the human-machine system. An additional important benefit is the prevention of negative impacts of sustained over- or underload on the mental health and cognitive capacity of the working population.

Acknowledgments. I would like to thank my student assistant Jon Scouten for daily operational support and proofreading. I would also like to express my sincere appreciation to Dr Gabriele Freude and Dr Uwe Rose for their valuable and constructive suggestions.

References

1. Brave, S., Nass, C.: Emotion in human-computer interaction. In: The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, pp. 81–96 (2002)
2. Debitz, U., Gruber, H., Richter, G.: Psychische Gesundheit am Arbeitsplatz. Teil 2: Erkennen, Beurteilen und Verhüten von Fehlbeanspruchungen, 3rd edn. InfoMedia Verlag (2003)
3. Hacker, W., Richter, P.: Psychische Fehlbeanspruchung. Psychische Ermüdung, Monotonie, Sättigung und Stress (Spezielle Arbeits- und Ingenieurpsychologie in Einzeldarstellungen), 2nd edn. Springer, Berlin (1984)
4. Gevins, A., Smith, M.E.: Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. Cerebral Cortex 10(9), 829–839 (2000)
5. Hagemann, K.: The alpha band as an electrophysiological indicator for internalized attention and high mental workload in real traffic driving. Ph.D. thesis, University of Düsseldorf, Germany (2008)
6. Hart, S.G., Staveland, L.E.: Development of the NASA TLX: results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, pp. 139–183. North Holland, Amsterdam (1988)
7. Kelly, S.P., Lalor, E.C., Reilly, R.B., Foxe, J.J.: Increases in Alpha Oscillatory Power Reflect an Active Retinotopic Mechanism for Distracter Suppression During Sustained Visuospatial Attention. Journal of Neurophysiology 95(6), 3844–3851 (2006), doi:10.1152/jn.01234.2005
8. Klimesch, W.: EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Research Reviews 29(2-3), 169–195 (1999)
9. Kompier, M.A.J., Kristensen, T.S.: Organisational work stress interventions in a theoretical, methodological and practical context. In: Dunham, J. (ed.) Stress in the Workplace: Past, Present and Future, pp. 164–190. Whurr Publishers, London (2001)
10. Landsbergis, P.A., Cahill, J., Schnall, P.: The changing organisation of work and the safety and health of working people: a commentary. Journal of Occupational Environmental Medicine 45(1), 61–72 (2003)
11. Lei, S.: Driver mental states monitoring based on brain signals. Ph.D. thesis, TU Berlin, Germany (2011)

12. May, J.F., Baldwin, C.L.: Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies. Transportation Research, Part F 12(2009), 218–224 (2008)
13. NIOSH - NORA Organization of work team members. The changing organization of work and the safety and health of working people. NIOSH-Publications Dissemination, Cincinnati (April 2002)
14. Parasuraman, R., Molloy, R., Singh, I.L.: Performance consequences of automation induced complacency. International Journal of Aviation Psychology 3(1), 1–23 (1993)
15. Parasuraman, R., Mouloua, M., Molloy, R.: Monitoring automation failures in human machine systems. In: Mouloua, M., Parasuraman, R. (eds.) Human Performance in Automated Systems: Current Research Trends, pp. 45–49. Earlbaum, Hilsdale (1994)
16. Posner, M.E., Peterson, S.E.: The attentional system of the human brain. Annual Review of Neuroscience 13, 25–42 (1990)
17. Sträter, O.: Warum passieren menschliche Fehler und was kann man dagegen tun?, Forum Prevention, AUVA - Allgemeine Unfallversicherungsanstalt, Wien (2001)
18. Unsworth, N., Heitz, R.P., Schrock, J.C., Engle, R.W.: An automated version of the operation span task. Behavior Research Methods 37, 498–505 (2005)

Neuronal Mechanisms of Working Memory Performance in Younger and Older Employees

Sergei A. Schapkin and Gabriele Freude

Federal Institute for Occupational Safety and Health, Berlin, Germany
schapkin.sergei@baua.bund.de, sschapkin@mail.ru

Abstract. As working memory (WM) is compromised with advancing age, older people may have performance deficits in WM tasks. This is probably due to a great number of WM operations which should be performed for extended periods of time. The reduction of a number of these operations was expected to reduce WM load and age-related deficits in WM performance. Fifty younger (29±3 years) and 49 older (55±3 years) healthy employees had to perform a visual 0-back (oddball) task and a 2-back task. Within the 2-back task, the short (3 or 4 items, low WM load) and long (5 or 6 items, high WM load) target-to-target sub-sequences were analysed separately. Older workers performed worse than younger ones at higher WM loads, except for the oddball condition and low WM load condition. The N2 latency of the event-related potentials (ERPs) increased with WM load and was generally longer in older than younger adults. In addition, the N2 latency decreased with WM load in younger adults but did not change in older ones. Older workers also showed a delayed P3a as well as a delayed and reduced P3b. By contrast, age-related enhancements of the occipital N1 and frontal P2 components under WM load were observed. The parietal slow positive wave (SPW) increased under high WM load but did not vary with age. The results indicate that older adults are able to compensate for age-related WM impairments when the amount of WM operations required does not exceed the limits of their WM capacity. The allocation of cognitive resources to stimulus encoding (N1) and memory retrieval (P2) are putative neuronal mechanisms for these WM improvements. However, older adults have maintenance problems at higher WM loads. This is associated with deficits in neuronal processes relating to response selection (N2), detection of changes in WM representations (P3a) and WM updating (P3b). These results provide a basis for the development of work load criteria and training opportunities for older workers who have to do complex work requiring working memory.

Keywords: aging, working memory, event-related potential.

1 Introduction

Working memory (WM), a system providing temporary storage and processing of information, is essential for flexible action regulation and adjustment to environmental demands [1]. WM decline with advancing age is well documented in the literature and is thought to be accounted for by progressive loss of neurons in the brain

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 70–81, 2014.

structures underlying WM [2]. The event-related potentials (ERPs) enable the examination of age differences in the allocation of processing resources to mental operations at different processing stages. The fronto-central N2 and parietal P3 (P3b) components are usually reduced and/or delayed in older people (see [3] for review), suggesting the existence of age-related deficits in executive processes associated with response monitoring [4] and working memory updating [5, 6]. The fronto-centrally distributed P3a component is closely related to the P3b and considered to be an index of attentional reorienting and novelty processing [3], [6]. However, the age effects on P3a in working memory tasks are not usually separated from those of P3b [7, 8, 9], which makes it difficult to disentangle the contribution of both processes to WM performance. The processes of early stimulus encoding also appear to be critical for WM [10]. The occipital N1 is thought to reflect selective amplification of sensory information to facilitate stimulus encoding [11], while a frontally distributed P2 component has been associated with top-down control over visual feature discrimination [12] or task-relevant stimulus evaluation [13] with both components tending to increase with advancing age [3]. However, the age-related changes in N1 and P2 are still matter of debate (see [3], for review). The parietal slow positive wave (SPW) is considered to be an index of effort and sustained attention, as this component increases with WM load [14], while age effects on the SPW are less well documented (e.g., [7]).

The n-back paradigm [15] is very well-suited to examine the key WM processes in real-time interaction while perceptual and motor demands are kept constant. Participants have to memorise a stimulus sequence (encoding, maintenance) and then decide whether a given stimulus matches one that appeared n trials ago (manipulation). This requires the continuous updating of memory representations (updating). A body of research using n-back tasks have demonstrated reliable age effects on performance and psychophysiological measures [7, 8, 9], [16].

Notably, the age-related performance decline in the n-back task has usually been inferred from data averaged across the whole trial block. Schmiedek et al. [16] used "lure" items embedded in a 2-back task and found that the age-related interference effects were only observed for lures up to four items back but not for longer trial subsequences. It follows that performance decline for short trial sub-sequences was predominantly due to lures and not to WM load per se. In other words, if no lures were presented, the age differences in performance would be small or absent. Based on the line of this reasoning, the present study analysed performance and ERPs for short and long target-to-target sub-sequences. In the low WM load condition participants had to process 3 to 4 items between two "neighbour" targets, while in the high WM load condition 5 to 6 items had to be processed. As processing of short sub-sequences requires a smaller number of WM operations than processing of long sub-sequences we expected that age differences at lower loads should be much smaller than at higher loads. The age-related increases of the N1 and P2 components were hypothesized to be an index of compensatory allocation of cognitive resources to early processing. The age-related frontal shift of the P3 is considered to be the compensatory activation of frontal brain areas to enhance cognitive control [7], [9]. With this in mind, the P3 frontality effect should be greater in older adults than in younger adults for both low and high WM load conditions than the oddball condition. We also assumed that

age-related reallocation of cognitive resources to early processing may deploy resources that are needed for the subsequent and more complex processes of response selection and memory updating. Hence, the delayed and/or reduced N2, P3a and P3b components should be observed in older compared with younger adults. The n-back task requires sustained attention for extended periods of time and older participants may compensate for WM deficits by enhanced on-task effort which may result in a SPW increase [7]. To examine this assumption age differences in the SPW were also analysed.

2 Method

2.1 Participants

Fifty younger (29±3 years) and 49 older (55±3 years) healthy employees were recruited through advertisements in local newspapers. The exclusion criteria were cardiovascular, neurological or psychiatric disorders, head injury, use of psychoactive medications or drugs. All participants were right-handed, native German speakers, had normal or corrected to normal vision, were currently employed with at least 20 hours per week, signed an informed consent and were compensated € 10 per hour.

2.2 Task

Twenty five different 12 x 18 mm Latin letters were presented successively in white on the black background for 200 ms each with an inter-stimulus interval of 1500 ms and a response window of max 1500 ms; each of them appeared with equal probability and was randomly distributed along the trial sequence. In the oddball task participants had to press a key with the right index finger when the letter "X" was displayed. In the 2-back task they had to maintain all incoming stimuli in memory and press a key if a letter was identical to the letter presented two trials previously. The oddball task consisted of 189 trials. The 2-back task consisted of 388 trials, where short (3 or 4 items) and long (5 or 6 items) target-to-target sub-sequences were analysed separately. The target probability (20%), physical and temporal features did not differ between the three conditions to avoid interference with WM load. The "neighbour" target-to-target sub-sequences consisted of letters which were highly different on perceptual features to avoid interference with lures [16]. The sub-sequences comprising 2, 3, 4, or 5 standards were quasi-randomly distributed within the oddball task and the 2-back task. Participants received training blocks in the 0-back task and the 2-back task until they attained 80% correct responses and thereafter conducted the main tasks.

2.3 EEG Recording

The EEG was continuously recorded from 24 electrodes (10-20 system) against Cz reference. The EOG was recorded from electrodes placed above and below the left

eye (vEOG) and next to the outer canti (hEOG). The signals were sampled and amplified with 2048 Hz (Brain Products LTD, Germany). Electrode impedance was kept below 10 kΩ. ERPs were re-referenced offline to linked mastoids. Eye movement artefacts were corrected using the Gratton & Coles algorithm. Epochs contained artefacts greater than ± 100 μV were excluded from analysis. The ERPs were filtered digitally with a 10 Hz low pass. The most prominent ERP components were identified by visual inspection of grand means at following sites: N1 at Oz, P2 at FCz, N2 at FCz, P3a at FCz, P3b at Pz, and SPW at Pz. The peak amplitudes and latencies at these sites were measured against 200 ms baseline in following time windows: N1 (100 - 150 ms), P2 (150 - 250 ms), N2 (200 - 300 ms), P3a (300 - 450 ms), P3b (350 - 700 ms), SPW (mean amplitude, 800 - 1200 ms).

2.4 Data Reduction and Statistical Analyses

All responses faster than 200 ms were excluded from the analysis. The correct RTs (raw and log-transformed) to targets, omission rates (OM), and false alarm rates (FA) were computed for the oddball task (baseline condition), short sub-sequences of the stimuli in the 2-back task (3 or 4 items, low WM load condition) and long sub-sequences of the stimuli in the 2-back task (5 or 6 items, high WM load condition). Similarly, the ERP components were also averaged for oddball, low load and high load conditions. Performance measures were subjected to an ANOVA with "Load" (oddball, low, high) as a within-subject factor and Age (younger, older) as a between-subject factor. The amplitudes and latencies of ERP components were subjected to an ANOVA with "Stimulus" (standard, target) and "Load" (oddball, low, high) as within-subject factors and Age (younger, older) as a between-subject factor. The Huynh-Feldt-corrected p-values were computed, if necessary. T-tests were applied to examine significant ANOVA effects. Statistical analyses were conducted using SPSS for Windows 18.0.

3 Results

3.1 Behavioural Data

The main effects of Load were significant on all performance measures (Fs > 109; ps < .001). Error rate was higher under high load (FA: 2.65%; OM: 10.53%) than low load (FA: .95%; OM: 7.25%) which in turn was higher than in the oddball condition (FA: .30%; OM: .28%). Surprisingly, RTs were longer at lower loads (558 ms) than higher loads (464ms) and the shortest in the oddball condition (357 ms). Correlations between speed and accuracy computed separately for each age group and load level did not reveal any speed-accuracy tradeoffs. Older adults performed as well as younger adults in the oddball condition (RT; FA; OM; younger: 358ms; .26%; .30%; older: 356ms; .34%; .25%) and low load condition (younger: 548 ms; .7%; 6.80%; older: 569 ms; 1.21%; 7.7%) but worse than younger adults at higher loads (younger: 449 ms; 2.15%; 8.51%; older: 479 ms; 3.15%; 12.56%) as expressed in an Age * Load interaction (ln RTs: $F(2, 194) = 3.59$, $p < .04$, $\eta^2 = .04$; false alarms: $F(2, 194) = 4.09$,

$p < .02$, $\eta^2 = .04$; omissions: $F(2, 194) = 3.98$, $p < .02$, $\eta^2 = .04$). Moreover, the Age * Load interaction on omission rate was due to the fact that the decline in accuracy going from the low load to the high load condition was significant for older adults but not for younger adults.

3.2 ERP Data

The ERP grand means as a function of age, load, and stimulus type are presented in the Figure 1.

N1. The N1 was larger to targets (-6.08 μV) than standards (-4.45 μV); Stimulus: $F(2,194) = 108.21$, $p < .001$, $\eta^2 = .53$. The N1 decreased under both low load (-4.93 μV) and high load (-5.00 μV) as compared to the oddball condition (-5.85 μV); Load: $F(2,194) = 19.01$, $p < .001$, $\eta^2 = .16$. This effect was seen for targets, while for standards the N1 decrease under low load compared with the oddball condition was found (Stimulus * Load: $F(2,194) = 43.45$, $p < .001$, $\eta^2 = .31$). Older adults showed a larger and later N1 (-6.15 μV, 136 ms) than younger ones (-4.37 μV, 127 ms) irrespective of stimulus type and load (Age, amplitude: $F(1, 97) = 5.58$, $p < .02$, $\eta^2 = .05$; latency: $F(1, 97) = 4.86$, $p < .03$, $\eta^2 = .05$).

P2. The P2 decreased in the low load condition (6.91 μV) as compared to both high load (7.45 μV) and oddball conditions (7.65 μV); Load: $F(2,194) = 4.86$, $p < .01$, $\eta^2 = .05$. This effect was seen for targets but not standards (Stimulus * Load: $F(2,194) = 15.96$, $p < .001$, $\eta^2 = .14$). The P2 latency decreased in both low load condition (167 ms) and high load condition (167 ms) relative to the oddball condition (174 ms), $F(2,194) = 14.05$, $p < .001$, $\eta^2 = .13$. A Load * Age interaction ($F(2,194) = 3.30$, $p < .05$, $\eta^2 = .03$) was attributed to a larger P2 in younger than older adults under both load conditions, while no age differences in the oddball task were found (low load, young: 6.01 μV; old: 7.81 μV; high load, young: 6.62 μV; old: 8.28 μV; oddball, young: 7.33 μV; old: 7.97 μV). Moreover, age groups exhibited a different P2 reactivity to high WM load: the P2 decreased in younger adults but not in older ones.

N2. The N2 was larger and later for standards (-.11 μV, 268 ms) than targets (2.29 μV, 246 ms); Stimulus, amplitude: $F(1, 97) = 64.33$, $p < .001$, $\eta^2 = .40$; latency: ($F(1, 97) = 89.58$, $p < .001$, $\eta^2 = .48$. A main effect of Load on the N2 amplitude ($F(2, 194) = 18.13$, $p < .001$, $\eta^2 = .16$) was due to a larger (i.e. less positive) N2 in the low load condition (-.02 μV) than both high load (1.66 μV) and oddball conditions (1.61 μV) with greater effect for targets than standards (Stimulus * Load: $F(2, 194) = 8.65$, $p < .001$, $\eta^2 = .08$). The N2 latency was longer in both oddball (260 ms) and low load (262 ms) conditions than in the high load condition (250 ms), $F(2, 194) = 12.27$, $p < .001$, $\eta^2 = .11$. The N2 latency decreased in the low load condition for targets (244 ms) compared with standards (280 ms) while no effects in other conditions were found (Stimulus * Load: $F(2, 194) = 10.73$, $p < .001$, $\eta^2 = .10$). A significant main effect of Age on the N2 latency ($F(1, 97) = 24.12$, $p < .001$, $\eta^2 = .19$) was due to a delayed N2 in older (268 ms) compared to younger adults (246 ms) irrespective of memory load. The N2 latency decreased at higher loads compared with both oddball and low load conditions in younger but not older adults (Load * Age: $F(2, 194) = 3.53$, $p < .03$, $\eta^2 = .04$.)

P3a. A main effect of Stimulus was due to a larger and earlier P3a to targets (10.76 μV, 389 ms) than standards (5.20 μV, 404 ms); amplitude: $F(1, 97) = 211.81$, $p < .001$, $\eta^2 = .69$; latency: $F(1, 97) = 4.18$, $p < .04$, $\eta^2 = .04$. A main effect of Load on the P3a amplitude ($F(2, 194) = 18.66$, $p < .001$, $\eta^2 = .16$) was attributed to a P3a reduction in the low load condition (6.63 μV) relative to both high load (8.30 μV) and oddball conditions (9.01 μV); $F(2, 194) = 25.65$, $p < .001$, $\eta^2 = .20$. The effect was more pronounced for targets than standards (Stimulus * Load: $F(2, 194) = 4.13$, $p < .02$, $\eta^2 = .04$). A longer P3a latency in both load conditions (low load: 407 ms, high load: 405 ms) than the oddball condition (376 ms) was found ($F(2, 194) = 18.93$, $p < .001$, $\eta^2 = .16$) with greater effect for targets than standards (Stimulus * Load: $F(2, 194) = 21.83$, $p < .001$, $\eta^2 = .18$). The P3a was delayed in older adults (420 ms) compared to younger adults (374 ms) irrespective of WM load ($F(1, 97) = 26.62$, $p < .001$, $\eta^2 = .21$).

P3b. As the P3b component was not reliably detectable in the standard-locked data, we analysed the P3b for targets only. The P3b was reduced and delayed under WM load (low load: 10.28 μV, 431 ms; high load: 12.69 μV, 449 ms) compared to the oddball condition (16.5 μV, 371 ms) as expressed in main effects of Load on the P3b amplitude ($F(2, 194) = 18.66$, $p < .001$, $\eta^2 = .16$) and latency ($F(2, 194) = 30.00$, $p < .001$, $\eta^2 = .24$). A reduced and delayed P3b for older adults (11.59 μV, 454 ms) compared to younger adults (14.73 μV, 408 ms) across conditions was also obtained (Age, amplitude*:* $F(1, 97) = 12.39$, $p < .001$, $\eta^2 = .12$; latency*:* $F(1, 97) = 11.36$, $p < .001$, $\eta^2 = .11$).

P3 Frontality. To test the "P3 frontality effect" in older adults, an ANOVA with the within-factors "Site" (FCz, Pz) and "Load" (oddball, low load, high load) for targets was performed. A significant "Site * Age" interaction ($F(1, 97) = 6.94$, $p < .01$, $\eta^2 = .07$) was due to age differences in the P3 amplitude at Pz (young: 14.78 μV, old: 11.57 μV) but not at FCz (young: 11.48 μV, old: 10.05 μV). The post-hoc tests revealed a parietal P3 maximum in younger adults irrespective of WM load (FCz vs. Pz, oddball: 13.19 μV, vs. 18.85 μV, low load: 9.62 μV vs. 11.55 μV, high load: 11.62 μV, vs. 13.93 μV, all ps < .001). By contrast, the P3 in older adults had the parietal maximum in the oddball condition but was more evenly distributed under WM load due to the P3 reduction at Pz (FCz vs. Pz, oddball: 10.90 μV vs. 18.85 μV, p <.001; low load: 8.31 μV vs. 9.03 μV, p < .16; high load: 10.96 μV vs. 11.48 μV, p < .27).

SPW. The SPW was larger for targets (3.44 μV) than standards (2.26 μV), (Stimulus: ($F(1, 97) = 24.76$, $p < .001$, $\eta^2 = .20$) and increased at higher WM loads (3.75 μV) compared to other conditions (oddball: 2.39 μV, low load: 2.42 μV); $F(2, 194) = 16.44$, $p < .001$, $\eta^2 = .14$. This increase was observed for targets but not standards (Stimulus * Load: $F(2, 194) = 14.20$, $p < .001$, $\eta^2 = .12$).

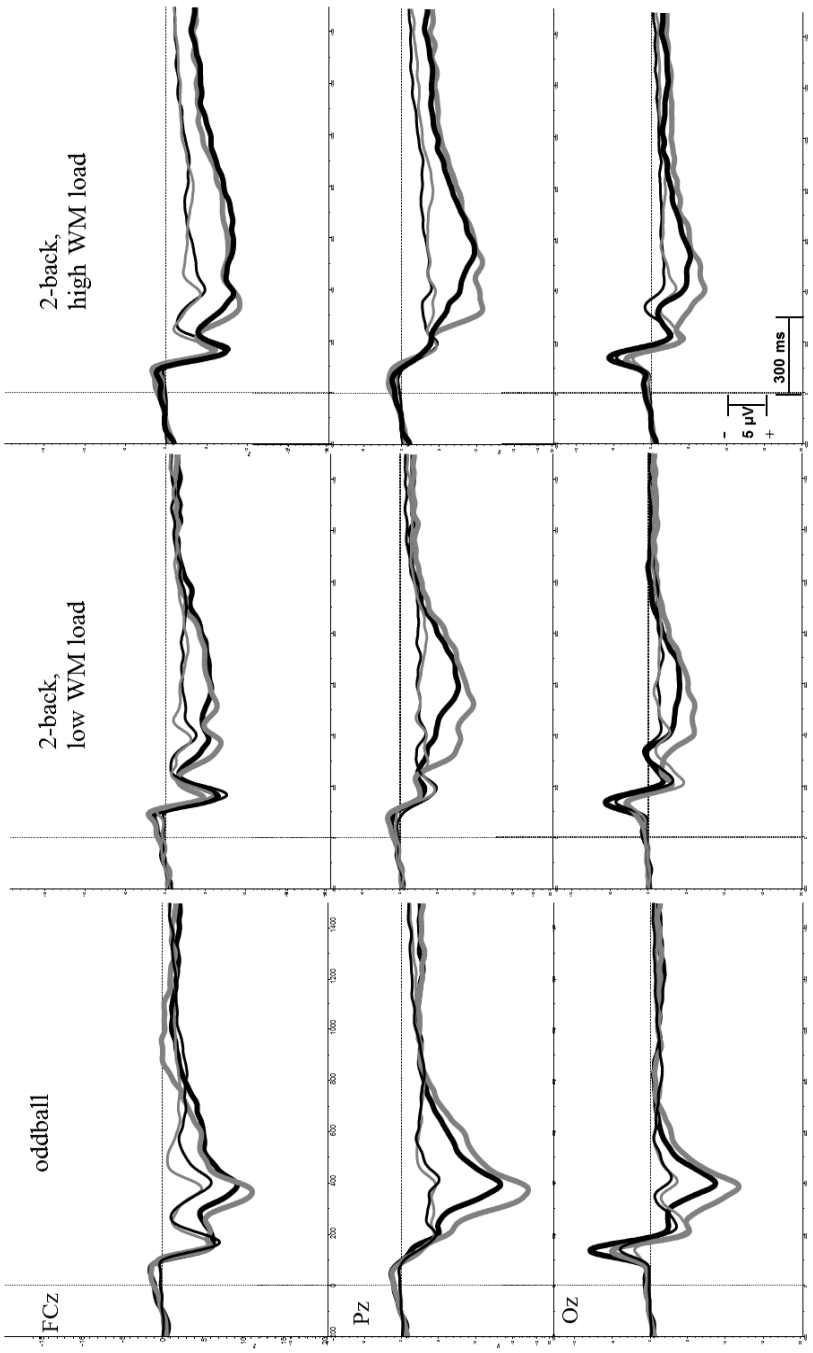


Fig.1. ERPs as function of age, working memory load and stimulus type. Black –old, gray –young, thick –targets, thin –standards.

Correlations between ERPs and Performance. To examine whether age groups differ in the involvement of cognitive processes during task performance, correlations between ERP components and performance measures in the high load condition were computed (Tab. 1). In younger participants, significant correlations between late ERP components and performance measures were found. Better performance (shorter RTs and/or lower error rate) was associated with larger P3a and P3b amplitudes and their shorter latencies to target stimuli. In addition, better performance was related to a reduced N2 and enhanced SPW for targets. Notably, significant correlations were found also for standards where better performance of younger adults was associated with both larger P3a and SPW as well as shorter P3a latency. By contrast, older adults revealed a smaller number of significant correlations between ERP components than younger adults. In older adults the P3a, P3b and SPW amplitudes negatively correlated with RTs while P3b latency positively correlated with omission rate.

Table 1. Pearson correlations between ERP and performance measures in the high WM load condition. RT – reaction time, % OM – omission percentage, % FA – false alarm percentage. Significances are in parentheses.

	Younger			Older		
Standards	RT	% OM	% FA	RT	% OM	% FA
N2 amplitude	-	-	-	-	-	-
P3a amplitude	-.35 (.01)	-	-	-	-	-
P3a latency	-	.36 (.01)	.39 (.005)	-	-	-
P3b amplitude	-	-	-	-	-	-
P3b latency	-	-	-	-	-	-
SPW amplitude	-	-.31 (.03)	-	-	-	-
Targets						
N2 amplitude	-.31 (.03)	-.28 (.05)	-.31 (.02)	-	-	-
P3a amplitude	-.40 (.004)	-	-	-.27 (.05)	-	-
P3a latency	.39 (.005)	.42 (.003)	.36 (.01)	-	-	-
P3b amplitude	-.29 (.03)	-	-	-.39 (.005)	-	-
P3b latency	.38 (.006)	.35 (.01)	.37 (.008)	-	.29 (.04)	-
SPW amplitude	-	-.37 (.008)	-.34 (.01)	-.31 (.03)	-	-

4 Discussion

4.1 Performance Data

The results of the present study largely agree with other literature demonstrating performance decline in older relative to younger adults in the 2-back task [7, 8, 9], [16]. Expanding on previous data we demonstrate that averaging performance in the 2-back task across the whole trial block is not precise enough to assess WM function, at least in this sample of healthy older employees. Nevertheless, analyses of the short and long target-to-target sub-sequences provide a more detailed evaluation of WM performance and putative neuronal mechanisms. The well-trained older individuals performed the 2-back task as well as younger ones when computational demands were low (3 or 4 items). By contrast, at higher computational demands (5 or 6 items), performance is observed to decline with age. The results confirm our assumption that older adults are able to compensate for performance deficits when the amount of WM operations required (e.g. encoding, maintenance, manipulation, updating) is rather small and items can be maintained in short-term memory (i.e. up to 6 seconds). The result is consistent with Cowan's model which postulates that humans are able actively maintain and process up to four items in WM [17].

4.2 ERP Data

Early Processing. Although the age-related increases of the N1 and P2 components have been obtained in different cognitive tasks [8], [10, 11, 12], their putative mechanisms are still unclear. It is known that the N1 and P2 are larger in attended than unattended stimuli and reflect early processes of stimulus encoding and evaluation respectively [11, 12]. Therefore, the age-related N1 increase we found suggests that older adults allocate more resources to stimulus encoding than younger ones irrespective of memory load. We also observed the N1 reduction in the 2-back task relative to the oddball task in both age groups. The result is commensurate with the 2-back task, where each standard stimulus must be assumed to be a possible target and thus processing resources should be shared between standards and targets. However, the N1 remained larger in older adults than in younger adults, suggesting less efficient allocation of processing resources in the former group. Notably, this age-related N1 increase in our study was also evident in the oddball task indicating that age differences in the N1 are unspecific to memory load but rather due to more general changes in executive control over stimulus encoding with advancing age [18].

The anterior P2 component is thought to reflect a top-down control over visual feature discrimination [12] or task-relevant stimulus evaluation [13]. Consistent with other studies [8] we found the P2 increase in older relative to younger adults. In addition to this, the age-related P2 increase was also observed under WM load except for the oddball condition. More importantly, the P2 decreased under WM load in younger adults but not in older adults. As the stimuli in the n-back task were easy to recognise, the P2 increase is unlikely to be a consequence of discrimination difficulty. Older participants were rather highly focused on early evaluation of whether a stimulus has

"targetness" properties, in order to select an appropriate response as quickly as possible. By contrast, the P2 reduction in younger participants may be due to the fact that the relevance evaluation proceeded more automatically and recruited fewer resources than in their older counterparts. Another explanation for the P2 increase may be memory retrieval processes which should be persistently active under WM in older adults. Conversely, younger adults may have relied less on retrieval mechanisms as expressed in P2 reduction under WM load. This interpretation is supported by our data from the memory-based switching task where participants had to retrieve stimulus-response (S-R) mappings from memory and apply them to a presented stimulus [19]. In similar fashion, the P2 in the switching task was larger in older than in younger adults. Moreover, the P2 latency positively correlated with RT mixing costs in older adults only suggesting extensive use of the retrieval mechanism to support maintenance of the S-R mappings in WM. Notably, the "retrieval" and "relevance evaluation" explanations for P2 are not mutually exclusive. A relevance evaluation may require extensive retrieval of items held in WM to match them with an incoming stimulus. The matching process is probably associated with the subsequent N2 component. Putting these interpretations together, we can surmise that the age effects on N1 and P2 components may be interpreted as compensatory allocation of processing resources for both stimulus encoding and evaluation via persistent recruitment of memory retrieval mechanisms in older adults.

Response Selection, Change Detection and Memory Updating. The functional role of the fronto-centrally distributed N2 component is still matter of debate [4], [20, 21, 22]. The N2 is elicited when participants are focusing on a stimulus to make a task-relevant decision [22] or have to match a current stimulus with WM representations [4]. The N2 is usually delayed when different S-R mappings should be held in WM [21]. In our n-back task we consider the N2 as an index of an executive control process which provides the priority of a relevant S-R mapping over irrelevant ones. In turn, the selected response inevitably requires changing and updating of WM content which probably elicit the following P3 component. Recent studies stressed a close relationship between the N2 and P3 in tasks requiring WM [19], [21]. Moreover, the shortening of the N2 latency with WM load that we found in younger adults suggests that their response selection is more efficient and/or less resource-consuming than in older adults.

Polich [6] proposed an integrative theory of P3a and P3b components which is well-suited for application to WM. He considers both components as indexes of different sub-processes constituting the common fronto-parietal network. An early attention process stemming from changes in working memory representations elicits the P3a. The attention-driven stimulus signal is then transmitted to parietal structures where WM updating occurs and the P3b is produced. All of this is consistent with the fact that in the present study, the P3a was seen for both standards and targets while the P3b was observed for targets only. It seems likely that the P3a reflects changes in WM representations elicited by each incoming stimulus. By contrast, the target stimuli require more elaborated processing and hence produce larger P3a and P3b. Significant correlations of P3b with reaction time and error rate in both age groups indicate the crucial role of continuous updating of information for WM performance. Younger

adults appear to use the mechanism of detection of changes in WM representations more extensively than older adults, as expressed in multiple correlations between performance and P3a for both targets and standards in the younger group. By contrast, only one significant correlation between performance and the P3a in older group was observed.

In agreement with other data ([7], [9]) we found a delayed P3a as well as a delayed and reduced P3b in older adults. The age-related frontal shift of the P3b is considered to be a compensatory activation of frontal brain mechanisms to enhance cognitive control [7]. However, in contrast to other studies using a similar paradigm ([7], [9]), we did not observe a larger frontal P3a in older as compared to younger adults (P3 frontality effect). Nevertheless, we found that the parietal P3 maximum in the oddball task changed to a more evenly distributed P3 under WM load in older adults while the parietal P3 maximum in younger adults persisted across conditions. Therefore, we did not obtain the age-related "P3 frontality effect" but rather a "P3b reduction effect" and cannot interpret the results as compensatory allocation of processing resources to frontal mechanisms.

One may ask which factors contribute to the updating deficits in older adults. Older participants might not be able to maintain their attention and effort during the task. However, the absence of age effects on SPW does not support this interpretation, as the SPW is considered to be an index of sustained attention and effort [14]. Other research has demonstrated that older participants are able to be engaged in demanding tasks and maintain on-task effort over time as well as their younger counterparts [20].

Recent studies suggest that reasons for the P3b reduction may originate in the processes preceding the P3b, namely in response selection. The age-related increase in the N2 latency we found agrees well with previous studies showing that difficulties in response selection may lead to a prolonged N2 which may overlap with the subsequent P3b and in turn reduce and prolong it [21]. Notably, the age-related lengthening of the N2, P3a, P3b latencies and the P3b reduction were already significant in the oddball task. The results indicate that key processes relying on WM weaken with age and this may already be observed in easy cognitive tasks where no WM load is imposed.

Another reason for the P3b reduction in older adults may be an inefficient allocation of cognitive resources. The age-related N1 and P2 increases found in the present study suggest that the compensatory allocation of cognitive resources to stimulus encoding and memory retrieval may result in the shortening of resources that are necessary for later operations like response selection (N2), detection of changes in WM representations (P3a) and WM updating (P3b). This explanation of age-related changes in neuronal mechanisms of working memory should be addressed in further studies.

References

1. Baddeley, A.: Working Memory. Science 255, 556–559 (1992)
2. Rodrigue, K.M., Kennedy, K.M.: The cognitive consequences of structural changes to the aging brain. In: Schaie, K.W., Willis, S.L. (eds.) Handbook of the Psychology of Aging, pp. 73–91. Elsevier, Amsterdam (2011)

3. Friedman, D.: The components of aging. In: Kappenman, E.S., Luck, S.J. (eds.) Oxford Handbook of Event-Related Potential Components. Oxford University Press, New York (2011)
4. Folstein, J.R., Van Petten, C.: Influence of cognitive control and mismatch on the N2 component of the ERP: a review. Psychophysiol. 45, 152–170 (2008)
5. Donchin, E., Coles, M.G.H.: Is the P300 component a manifestation of context updating? Behav. Brain Sci. 11, 355–425 (1988)
6. Polich, J.: Updating P300: An integrative theory of P3a and P3b. Clin. Neurophysiol. 118, 2128–2148 (2007)
7. Daffner, K.R., Chong, H., Sun, X., Tarbi, E.C., Riis, J.L., McGinnis, S.M., Holcomb, P.J.: Mechanisms underlying age- and performance-related differences in working memory. J. Cogn. Neurosci. 23, 1298–1314 (2011)
8. McEvoy, L.K., Pellouchoud, E., Smith, M.E., Gevins, A.: Neurophysiological signals of working memory in normal aging. Cogn. Brain Res. 11, 363–376 (2001)
9. Wild-Wall, N., Falkenstein, M., Gajewski, P.D.: Age-related differences in working memory performance in a 2-back task. Front. Psychol. 2, 186 (2011)
10. Gazzaley, A., Clapp, W., Kelley, J., McEvoy, K., Knight, R.T., D'Esposito, M.: Age-related top-down suppression deficit in the early stages of cortical visual memory processing. PNAS 105, 13122–13126 (2008)
11. Hillyard, S.A., Anllo-Vento, L.: Event-related brain potential in the study of visual selective attention. PANAS 95, 781–785 (1998)
12. Luck, S.J., Hillyard, S.A.: Electrophysiological correlates of feature analysis during visual search. Psychophysiol. 31, 291–308 (1994)
13. Potts, G.F.: An ERP index of task relevance evaluation of visual stimuli. Brain and Cogn. 56, 5–13 (2004)
14. Ruchkin, D.S., Johnson Jr., R., Canoune, H., Ritter, W.: Short-term memory storage and retention: an event-related brain potential study. Electroencephalogr. Clin. Neurophysiol. 76, 419–439 (1990)
15. Gevins, A., Smith, M.E., McEvoy, L., Yu, D.: High-resolution, E.E.G.: EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. Cereb. Cortex 7, 374–385 (1997)
16. Schmiedek, F., Li, S.C., Lindenberger, U.: Interference and facilitation in spatial working memory: age-associated differences in lure effects in the n-back paradigm. Psychol. Aging 24, 203–210 (2009)
17. Cowan, N.: The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behav. Brain Sci. 24, 87–185 (2001)
18. Chao, L.L., Knight, R.T.: Prefrontal deficits in attention and inhibitory control with aging. Cereb. Cortex 7, 63–69 (1997)
19. Schapkin, S.A., Gajewski, P.D., Freude, G.: Age differences in memory-based task switching with and without cues: An ERP study. Journal of Psychophysiology (in press, 2014)
20. Falkenstein, M., Hoormann, J., Hohnsbein, J.: Inhibition-related ERP components: variation with modality, age, and time on-task. J. Psychophysiol. 16, 167–175 (2002)
21. Gajewski, P.D., Falkenstein, M.: Diversity of the P3 in the task-switching paradigm. Brain Research 1411, 87–97 (2011)
22. Ritter, W., Simson, R., Vaughan Jr., H.G., Friedman, D.: A brain event related to the making of a sensory discrimination. Science 203, 1358–1361 (1979)

A Method to Reveal Workload Weak-Resilience-Signals at a Rail Control Post

Aron W. Siegel[1] and Jan Maarten Schraagen[1,2]

[1] University of Twente/GW, P.O. Box 217, 7500 AE Enschede, The Netherlands
{A.W.Siegel,J.M.C.Schraagen}@UTwente.nl
[2] TNO Earth, Life, and Social Sciences, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
Jan_Maarten.Schraagen@TNO.nl

Abstract. Reorganization of a rail control post may affect its ability to cope with unexpected disruptions. The term 'resilience', the ability to manage spare adaptive capacity when unexpected events occur, encapsulates this situation. This paper focuses on the workload adaptive capacity through a method for revealing workload weak-resilience-signals (WRS). Three different workload measurements are adapted to identify structural changes in workload. The first, executed cognitive task load, targets system activities. The second, integrated workload scale, is a subjective measure. The last, heart rate variability, identifies physiological arousal because of workload. An experiment is designed to identify the workload change and distribution across group members during disruptions. A newly defined Stretch, the reaction of the system to an external cluster-event, is used to reveal a workload WRS. The method is suitable for real-time usage and provides the means for the rail signaler to influence the system through his subjective workload perception.

Keywords: Resilience, weak resilience signal, WRS, objective and subjective Stretch, workload, rail operations, rail control post.

1 Introduction

Organizations restructure to improve their work efficiency. This efficiency step can, however, affect their spare, and sometimes hidden, adaptive capacity needed when an unexpected disruption occurs. In addition, this efficiency step can also affect the organization's ability to manage this capacity. A socio-technical organization needs this ability to cope with disruptions, commonly referred to as 'resilience' [1]. As improved work efficiency may conflict with an organization's resilience due to common resource demands, we need methods to identify this potential conflict. This paper deals with such a method and concentrates on the restructuring of a rail control post. A rail control post is responsible for a large area containing railway stations, controlled by rail signalers managing the traffic on the rail infrastructure. The post is 24/7 active with between 10 to 20 rail professionals, depending on the number of railway stations covered. A rail control post is an example of a socio-technical system due to the critical human-system interaction. Siegel & Schraagen [2] argued

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 82–93, 2014.

that resilience in rail operations influences the rail system's operating state in three main areas: safety, performance (capacity and punctuality), and workload. In each of these areas, a weak resilience signal (WRS) may occur indicating a possible change of the system resilience, which needs further investigation to draw a solid conclusion. In this study, we focus on one area – workload. Changes in workload due to an organizational restructuring imply a change of the workload capacity needed during disruptions. This change is a reflection of a workload weak resilience signal (WRS), assuming that a decrease in the workload capacity lowers the ability of a socio-technical system to cope with disruptions and calamities, thus decreasing its resilience. The method described in this paper provides a means for investigating this assumption. Specifically, it aims at answering the following research questions: 1) Can workload measurements identify the human consequences of an organizational change, and 2) does such a change imply a possible impact on the resilience of the system? The activities are performed in a real operational environment where improvement or degradation may occur. Although a WRS is intended to signal a degradation, an improvement will be sufficient as well for demonstrating the concept, which will be relevant for a reversed restructuring.

In the following sections, we describe the setting of a rail control post when it is restructuring tasks among the rail signalers and follow this description with the approach of the method used. Afterwards, we elaborate on the method ingredients and their specifics about the setting. We finish the paper describing an experiment design and analysis approach, and a discussion.

2 Setting

The setting is a rail control post with m_{Post} workstations and n_{Post} rail signalers evaluating a new organization form to increase their performance. Each workstation, WS_j, is allocated to a set of railway stations and operated by one rail signaler, RS_i, who is responsible for all the workstations' aspects. These aspects are roughly divided into logistics and safety. The workstations are split into two groups. The first group, G_T, is the target group that will reorganize to improve its performance. The second group, G_R, is the reference group that will not reorganize throughout the testing period. All the n_{Post} rail signalers of the control post may be allocated to each of the groups and to each of its workstations. In group G_T there are m_T workstations WS_{Tj} and in group G_R there are m_R workstations WS_{Rj}. In addition, there is a calamity workstation, WS_{cal}, which is added to give support to the workstation being at the core of a calamity. The calamity workstation can be added to each group, G_T or G_R. The setting is depicted in Fig. 1:

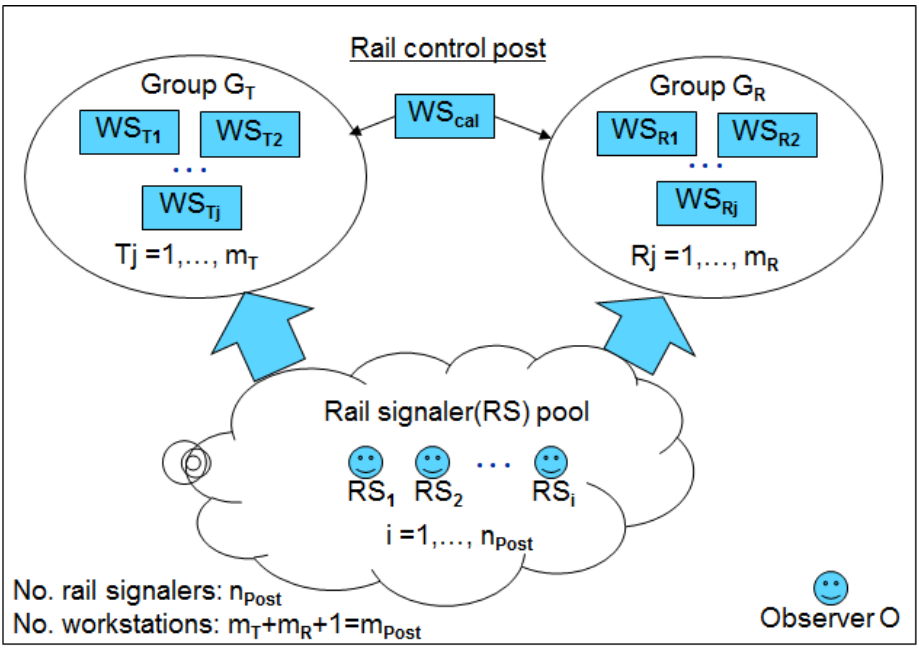


Fig. 1. Rail control-post setting with observer O

3 Approach

Pickup, Wilson, Nichols, & Smith [3] developed a conceptual framework of mental workload for railway signallers and differentiate among three types of mental load: (1) Imposed load, through the task characteristics; (2) Internal load and (3) perceived load, through the individual characteristics of the controller. We suggest using three different measurements to be able, to some extent, to differentiate among the influences of the mental load types: 1) external cognitive task load (XTL), 2) subjective workload, 3) physiological arousal created by workload. This is in line as well with Veltman [4] who argued that one needs performance data, subjective data and physiological data for a complete understanding of workload. Neerincx [5] modelled cognitive task load (CTL) in three dimensions: task complexity, task duration and task switching. We build upon this theory to compose XTL. We use the Integrated Workload Scale (IWS) [6] to measure subjective load and the extensively-researched heart rate variability (HRV) to identify physiological arousal due to workload change [7–12]. In the next sections, we elaborate these measures in more detail for the setting described above.

3.1 External Cognitive Task Load (XTL)

Rail signalers' task execution can be divided into four main activities, which are measurable within the system: 1) monitoring (Mon), 2) plan mutations (Plan), 3)

manual actions (Man), and 4) communication (Com). We assume that monitoring is in proportion with automated activities executed by the system. This assumption refers to imposed task load, while in reality the rail controller can actually ignore it. Monitoring is measured by counting all the automated activities. These activities are counted in 5 minute base-slots, used throughout all the types of measurement for ease of comparison. We normalize these counts by dividing them by the maximum count (Mon_{max}) occurred throughout the test period. This causes the measurement to be normalized between 0 and 1. This same idea is applied to normalizing the plan mutations and the manual actions. Each of them are counted within the 5 minute base-slot and divided by the maximum count, $Plan_{max}$ and Man_{max} respectively, throughout the test period. The communication normalization is done differently. Communication is defined by the percentage of verbal exchanges over the phone, which is measureable, during the 5 minute base-slot. If the XTL is concerning a group, then 100% communication is defined by all members talking the whole 5 minutes.

The combination of these four normalized activities refers to task complexity as stated by Neerincx [5]. However, Neerincx used the Skill-Rule-Knowledge (SRK) model [13] to express task complexity by rating each task on its SRK cognition load level. We have chosen to describe the cognitive load of each of the four activities and track their identity throughout the whole process. Monitoring is about following the automated system. Updating the planning is a logistics task and coordination task with external parties, such as the train operators. Manual activities include direct operations on the infrastructure instead of the automated system. This demands a logistic understanding as well and needs good perception and insight of the infrastructure in the field. Telephone conversation has a large cognitive task load. In most cases, the signaler needs to understand the logistic and infrastructure situation outside while talking with the person on the phone, such as a train driver, and visualize the issues in the field. It is challenging to perform another task during demanding telephone conversations. In addition to these activities, task switching and task duration are two extra dimensions amplifying the workload. To estimate the number of task switches, we look at the task activations and count them in each time slot as long as they are activated, to reflect the task duration. In figure 2, we list the task activations imposed on a particular workstation or on the group R or T. These activations result in the activities discussed above and result in workload we are measuring by XTL, IWS and HRV. We divide the number of activations, occurring in the 5 minute base-slot, by the maximum activations occurring throughout the test period to achieve a normalized switching factor between 0 and 1. Task switching and duration are a cognitive add-on to the activity load. With the same activity load, 0 to n parallel task switches can occur, behaving like a cognitive amplifier to the activity load. We add one to the normalized switching factor to act as a cognitive amplifier by becoming a growth multiplier of the activity load. Graphically, the multiplication will show jumps attracting the attention needed for interpretation. Thus, the switching factor becomes:

$$K_{switch} = \frac{number\ of\ activations\ in\ 5\ \mathrm{min}\ base\text{-}slot}{maximum\ number\ of\ activations\ in\ 5\ \mathrm{min}\ base\text{-}slot} + 1 \quad (1)$$

We multiply the task switching factor with the added four normalized tasks to achieve a combined XTL number. This approach will create a number between 0 and 8 to be used as an overall graphical indication on the XTL magnitude and change. Maximum load due to task execution is: 4 X 1 = 4, multiplied by a maximum switching factor: 2 X 4 = 8. However, it is important to present all the components and their relationships separately, to understand the situation. We will discuss this further in the "Experiment Design and Analysis Approach" section.

The XTL calculations can be performed for the following units:

- Workstation WS_j; $XTL_{WS\text{-}j}$, j=1,…, m_{Post}

$$XTL_{WS\text{-}j} = K_{switch\text{-}WS\text{-}j} \times \left(\frac{Mon_{WS\text{-}j}}{Mon_{max}} + \frac{Plan_{WS\text{-}j}}{Plan_{max}} + \frac{Man_{WS\text{-}j}}{Man_{max}} + Com_{WS\text{-}j} \right) \quad (2)$$

- Target group; $XTL_{G\text{-}T}$ for WS_{Tj}, Tj=1,…,m_T + 1 (WS_{CAL})

$$XTL_{G\text{-}T} = \frac{1}{m_T + 1} \times \sum_{\mathrm{Tj}=1}^{m_T+1} XTL_{WS\text{-}Tj} \quad (3)$$

- Reference group; $XTL_{G\text{-}R}$ for WS_{Rj}, Rj=1,…,m_R + 1 (WS_{CAL})

$$XTL_{G\text{-}R} = \frac{1}{m_R + 1} \times \sum_{\mathrm{Rj}=1}^{m_R+1} XTL_{WS\text{-}Rj} \quad (4)$$

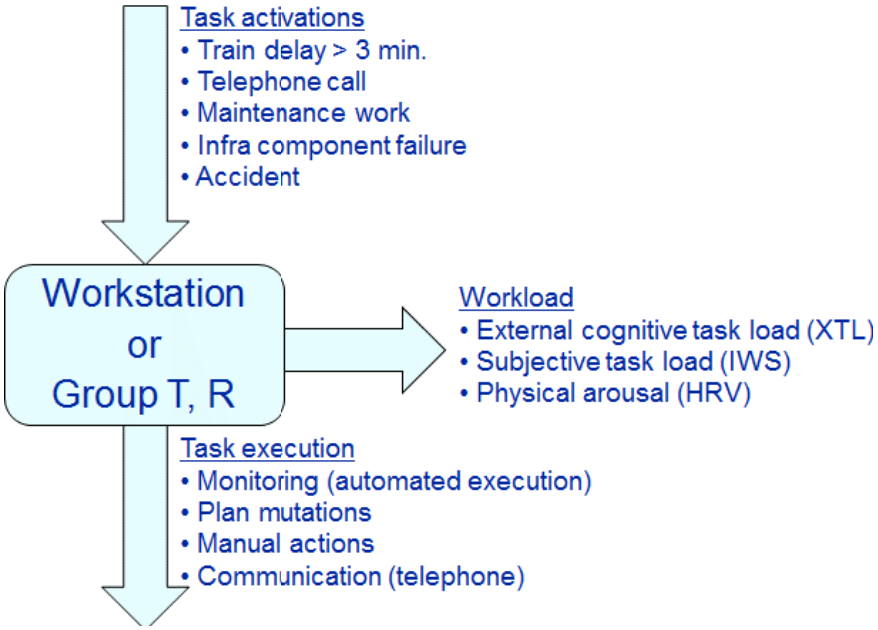


Fig. 2. Task flow per workstation or group

3.2 Integrated Workload Scale (IWS)

The Integrated Workload Scale [6] for a computer program runs on a laptop near each work station. The rail signaler RS_i, working at work station WS_j, is alerted every 5 minutes by a peripheral blinking rectangle, to rate his or her subjective workload. He or she is presented a 9 scale figure with the following text (in Dutch) (see figure 3):

1. Not demanding
2. Minimal effort
3. Some spare time
4. Moderate effort
5. Moderate pressure
6. Very busy
7. Extreme effort
8. Struggling to keep up
9. Work too demanding

The rail signaler has the possibility to add a comment to his or her rating and gets a graphic overview of his or her scoring (see Fig. 3).

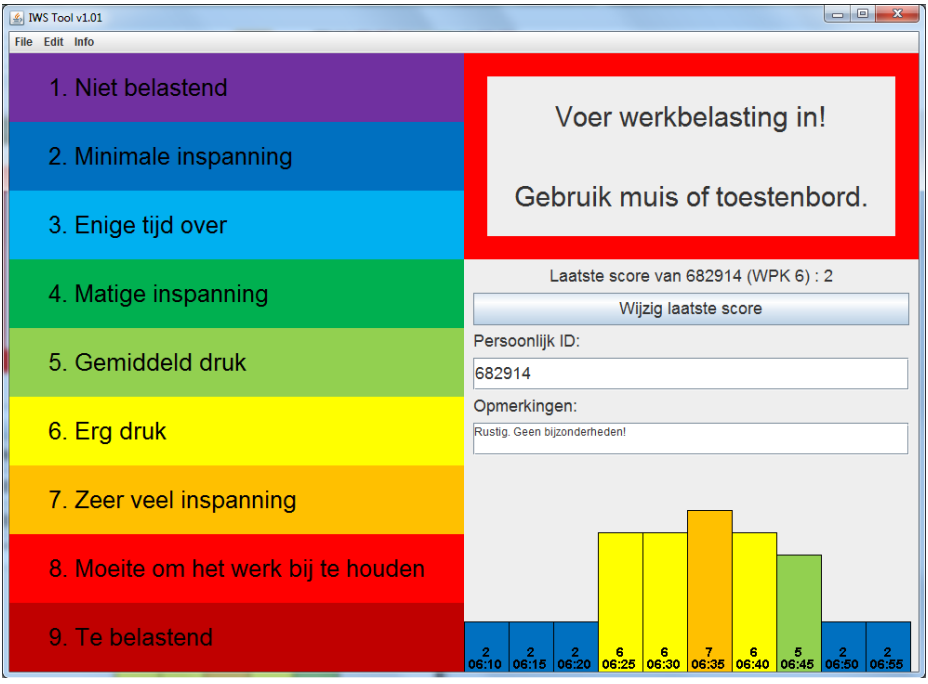


Fig. 3. IWS application screenshot (upper-right red rectangle blinks to draw attention)

The $IWS_{WS\text{-}j,RS\text{-}i}$ is initially rated personally by rail signaler RS_i. We calibrate their scoring in order to combine it with the scoring of other signalers. Wilms & Zeilstra [14] have calibrated only with a quiet period rather than a rush hour situation. We

propose to extend the calibration to two situations: the quiet period and rush hour. Both situations are well defined by the normal planning and therefore suitable for calibration. Each signaler is asked about their rating when nothing special is happening between 10-11 AM, defining a quiet period, and between 5-6 PM, defining rush hour. We perform a linear transformation to the IWS calibrated values 2 and 5, matching the two situations, while maximizing it to 9. This results in a calibrated $IWS^{C}_{WS\text{-}j}$ for every 5 minute base-slot (see figure 4).

The IWS calculated for each group is the average of all the workstations within the group. For the target group GT: IWS^{C}_{GT} = average($IWS_{WS\text{-}j}$, j=1, .., mT), and for the reference group GR: IWS^{C}_{GR} = average($IWS_{WS\text{-}j}$, j=1, .., mR).

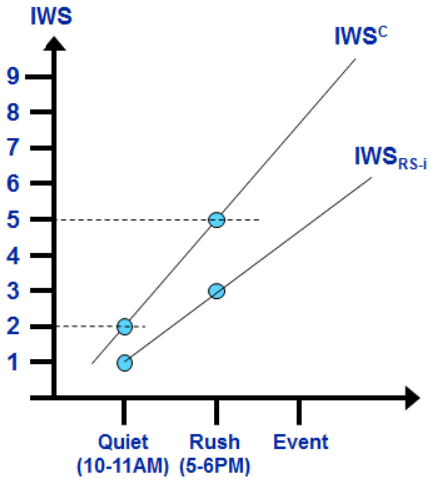


Fig. 4. IWS calibration

After calibration, it is possible to plot IWS together with XTL against the time for each workstation and group, since IWS has the same 5 minute base-slot as the XTL. We will use this relation in the "Experiment Design and Analysis Approach" section.

During the experiment an independent observer uses the IWS tool to rate $IWS_{O,G\text{-}T}$ for the whole target group and $IWS_{O,G\text{-}R}$ for the whole reference group. The observer uses the comment box to record relevant events.

3.3 Heart Rate Variability (HRV)

The heart rate variability is measured with a heartbeat device (Zephyr HxM BT), which is positioned on a breast strap and transfers its data to a laptop near each workstation. Every signaler wears the heartbeat device at the start of their work. The device sends continuous strings with recorded R-R beats in msec. The R-R interval is the time between subsequent R peaks of an ECG waveform. The information is stored on the laptop and post-processed the following day. HRV can be calculated in various ways roughly divided in time domain methods and frequency domain methods [11].

We use the most common occupational health method [12]: SDNN, the standard deviation (SD) of all normal-to-normal (NN) intervals, from the time domain and the low-high frequency (LF/HF) ratio in the frequency domain. We calculate both measures in the same 5 min base-slot used for the calculation of XTL and IWS. The SDNN is calculated by the standard deviation of R-R, the peak to peak interval, which is a very close measure of N-N, normal to normal interval. The LF/HF ratio is calculated through a discrete Fourier transform (DFT) of the first 256 measures, imposed by the DFT methodology using 2^n samples, in the 5 min base-slot. For a heartbeat rate of 80 bpm there are about 400 R-R samples in 5 minutes implying 256 measures to be the maximum integer power of two. The LF is the spectral integral of frequencies between 0.04 Hz and 0.15 Hz. The HF uses frequencies between 0.15 Hz and 0.4 Hz. We use these two HRV measures in the analysis.

4 Experiment Design and Analysis Approach

The control post, searching to optimize its processes, is restructuring around corridors. Until recently, all rail signalers were allocated individually to a few railway stations, being responsible for safety and the logistics through planning. This way of working is typical for the reference group. The target group, around one corridor, will divide safety and logistics responsibilities differently. One rail signaler will be responsible for all the planning activities within the corridor and the other rail signalers will only deal with safety. The experiment is designed to have two measurement periods of one week (Monday-Friday). The first period is a baseline measurement when no organizational changes have yet taken place. The second period is at least one month after the target group has reorganized and settled into the new setting. The plan is to record XTL, IWS and HRV 24 hours a day, but can be less due to practical reasons. Phenomena occurring in the target group before and after the change are likely to be caused by the organizational change but may also be caused by the measurements themselves [15]. We ignore the last possibility, since we are, for practical reasons, not able to perform an extra measurement set without a reorganization to show the measurement influence. Under these conditions we assume that phenomena, which do not occur in the reference group are due solely to the reorganization of the target group.

The analysis for each measurement period focuses at first separately on each of the workload methods: XTL, IWS and HRV. The external cognitive task load (XTL) with its 5 rail components – monitoring, planning, manual actions, communication and parallel tasking - is the main basis of estimating the workload. The XTL is the main basis since it is objectively measurable and represents facts derived from the system, while IWS and HRV have a more subjective character. The XTL information will be organized for each workstation in the 5 minute base-slots t5: $XTL_{WS\text{-}j}(t5)$, $j=1,\ldots,m_{post}$. Afterwards, XTL is clustered for the groups T and R in the same base-slots: $XTL_{G\text{-}T}(t5)$, $XTL_{G\text{-}R}(t5)$. These XTL values are plotted against the time (see Fig. 5). The IWS is calibrated for each person and combined for each work station as well and plotted together with XTL, but with its own y-axis (see Fig. 5). IWS and

XTL behave differently. The XTL is more steady due to human estimation which changes gradually. The XTL is derived from system parameters, which causes a more wavering character. In order to relate the IWS and XTL measurements, a new term is introduced – Stretch. A Stretch is the accumulative workload effort during a period initially defined by IWS rising from a baseline until it returns to the baseline. The IWS-baseline is defined as the steady state IWS rating before and after a disruption. However, the activity in the system may have started earlier and ended later. Therefore, the starting moment of a Stretch is adjusted to the first XTL-minimum moment before the IWS rising. Similarly, the ending moment of a Stretch is adjusted to the first XTL-minimum moment after the IWS return. In other words, a Stretch is the reaction of the system to an external cluster-event. We use the term cluster-event, since more than one event may occur during a stretch. An Objective Stretch is the name of the area under XTL, since it is objectively measured. We name the area under IWS a Subjective Stretch due to its subjective IWS rating. The division of the Subjective Stretch by the Objective Stretch is called the Stretch ratio and is used to identify changes in the workload revealing a workload WRS. An average larger Stretch ratio during the period after the reorganization compared to the baseline period, indicates more subjective workload on similar external events. The Objective Stretch is used to identify an absolute workload growth, throughout a specific period like a day or a workweek. In the example of Fig. 6 we have plotted the Stretch ratio during a week for groups T & R. A significant change in the Stretch ratio, comparing two weeks, indicates a change in the relative workload and may be considered as a workload WRS.

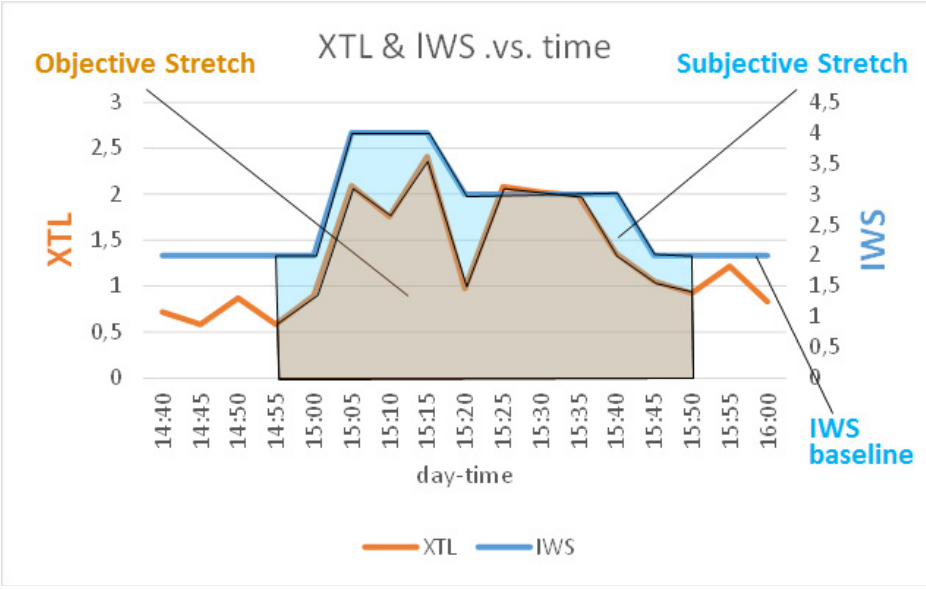


Fig. 5. Objective and Subjective Stretch

We use the HRV to validate the subjective IWS ratings. We expect the HRV to decline during a stretch as evidence that a growing IWS is an expression of a growing mental load.[12]

The XTL components are used to show the workload distribution among the members of each group R and T. The standard deviation (SD) presents the work distribution for each of the activities (for an example, see Fig. 6). The work distribution can be calculated for each stretch or for longer periods like a day or week and is calculated for each Group. The work distribution among group members, can help to explain the reasoning behind structural changes of Stretches and amplify the presence of a workload WRS.

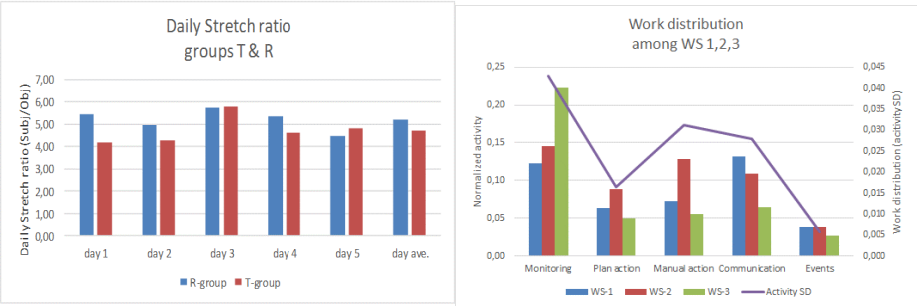


Fig. 6. Analysis examples: Daily Stretch ratio and work distribution

5 Discussion

The method described in this paper will be used for the experiment design worked out in the previous section. It provides the means to analyze the relation between workload and resilience, in particular, the method enables research on weak resilience signals (WRS), as explained in the introduction. The underlying assumption is that a decrease in the workload capacity lowers the ability of a socio-technical system to cope with disruptions and calamities. The decrease in workload is identified through Stretches, the reaction of the system to an external cluster-event, which has three types: objective, subjective and their ratio. The work distribution analysis ability provides deeper reasoning on Stretch changes. In future research, we plan to compare these results with other methods of identifying the resilience state of a team, such as questionnaires. This research should underpin in more depth the relation between workload and resilience.

Our method is applicable beyond the designed experiment and opens new dimensions for incorporating the human factor in decisions and control throughout disruption handling. The method has resulted in a real-time software tool, providing the possibility for the rail signaler to influence operations, through his or her subjective workload perception combined with objective measurements. We choose measurement variables that are available in real-time. This allows the possibility to provide feedback on the resilience state during operations and stimulates corrections

on the spot. In addition, we have designed the workload measurement balancing the objective and subjective. The rail signaler can express his or her personal opinion, show that to his or her environment and directly influence decisions and actions. We present the personal opinion in relation to the objectively measured workload and physiological arousal, providing a balanced view. Leveling the human state with the technical one is a change from today's situation in the rail organization, where the responsibility for human well-being is hidden in the lower level management on the work floor. By using this software tooling, it provides real-time insight of the human status to all levels in the organization. The impact of this tooling and methodology on the system resilience needs further research as well.

Acknowledgement. We thank Jaldert van der Werf for his development of the IWS and analysis software tooling, and his contribution to the experiment. We appreciate the constructive comments of the reviewers and thank Alfons Schaafsma guidance. This research was conducted within the RAILROAD project and is supported by ProRail and the Netherlands organization for scientific research (NWO) (under grant 438-12-306).

References

1. Hollnagel, E., Woods, D.D., Leveson, N. (eds.): Resilience engineering: concepts and percepts. Ashgate Publishing Limited, Hampshire (2006)
2. Siegel, A.W., Schraagen, J.M.: Developing resilience signals for the Dutch railway system. In: 5th Resilience Engineering Symposium (in press)
3. Pickup, L., Wilson, J.R., Nichols, S., Smith, S.: A conceptual framework of mental workload and the development of a self-supporting integrated workoad scale for railway signallers. In: Wilson, J., Norris, B.J., Clarke, T., Mills, A. (eds.) Rail Human Factors, pp. 319–329. Ashgate, Surrey (2005)
4. Veltman, J.A., Gaillard, A.W.K.: Pilot workload evaluated with subjective and physiological measures. In: Brookhuis, K., Weikert, C., Moraal, J., de Waard, D. (eds.) Aging and Human Factors, pp. 107–128. University of Groningen, Haren (1996)
5. Neerincx, M.A.: Cognitive task load analysis: allocating tasks and designing support. In: Hollnagel, E. (ed.) Handbook of Cognitive Task Design, pp. 283–305. Lawrence Erlbaum Associates, Mahwah (2003)
6. 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, 681–693 (2005)
7. Billman, G.E.: Heart rate variability - a historical perspective. Front. Physiol. 2, 86 (2011)
8. Goedhart, A.D., van der Sluis, S., Houtveen, J.H., Willemsen, G., de Geus, E.J.C.: Comparison of time and frequency domain measures of RSA in ambulatory recordings. Psychophysiology 44, 203–215 (2007)
9. Hoover, A., Singh, A., Fishel-Brown, S., Muth, E.: Real-time detection of workload changes using heart rate variability. Biomed. Signal Process. Control 7, 333–341 (2012)
10. Jorna, P.G.A.M.: Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. Biol. Psychol. 34, 237–257 (1992)

11. Malik, M.: Heart Rate Variability. Ann. Noninvasive Electrocardiol. 1, 151–181 (1996)
12. Togo, F., Takahashi, M.: Heart rate variability in occupational health-a systematic review. Ind. Health 47, 589–602 (2009)
13. Rasmussen, J.: Risk management in a dynamic society: a modelling problem. Saf. Sci. 27, 183–213 (1997)
14. Wilms, M.S., Zeilstra, M.P.: Subjective mental workload of Dutch train dispatchers: Validation of IWS in a practical setting. In: 4th International Conference on Rail Human Factor, pp. 641–650 (2013)
15. Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin Company, Boston (2002)

An Analysis of Pilot's Physiological Reactions in Different Flight Phases

Zhen Wang and Shan Fu

School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai, 20040, P.R. China
b2wz@sjtu.edu.cn

Abstract. Human errors have become the major threat to flight safety. Improper workload imposed on the pilot was supposed to be one of the most critical causes for human error. Among the various workload measure techniques, physiological measures were promising because of its objectivity and capability of continuous measure. However, the mechanisms of the physiological reactions were complex. The sensitivity and diagnosticity of the physiological parameters should be carefully discussed. This paper study several physiological reactions in a simulated flight task including blink, saccade, fixation, pupil diameter, heart rate, respiration etc. Statistical analysis was made to test the sensitivities of the physiological parameters and the diagnosticity of parameter was also discussed in the paper. These results could provide some guidance for the selection of the physiological parameters during the assessment of the pilot's workload.

Keywords: flight safety, workload, simulated flight, physiological measures.

1 Introduction

Several recent flight accidents reminded us that flight safety was always the most concerning problem in aviation. In the last decades the reliabilities of the aircraft and the onboard systems had been largely improved. More and more evidences showed that human errors have become the major threat to the flight safety[1].

Many researchers agreed that the improper workload imposed on the human operators is one of the most critical causes of human errors[2, 3]. However, workload is the experience of the human operator, it cannot be measured directly. And it is even hard to give it a clear description because workload is a multidimensional construct[4, 5]. Different techniques have been developed to measure workload. Performance measures, subjective rating scale and physiological measures are the most representative workload measure techniques[6, 7]. Yet, the performance measures had been proved to be inadequate to describe workload, this could be easily explained by that human operator could maintain the performance by paying more efforts[2]. Though subjective rating scales are the most frequently used workload measure technique, they still have some shortcomings. For example, they were depending on the operator's memories and cannot be executed continuously in the task[6]. Thus

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 94–103, 2014.

they could not provide a detailed evaluation of the operator's workload. Physiological measures are promising workload measure techniques because the physiological parameters could be measured continuously in the whole process of the task[8] and moreover, several physiological parameters have been shown to be sensitive indicators in some particular tasks[9, 10]. However, the mechanisms of the physiological reactions were complex, thus the sensitivity of different physiological parameters to the workload should be discussed carefully.

In this study, several physiological parameters were measured in a simulated flight task. The data were analyzed to find out the sensitivities of the physiological parameters to the different level of pilot's workload. In the rest of this paper, the experiment would be described firstly; and the experimental results would be presented; then statistical analysis was executed to study the sensitivity of the data. A discussion about the results would be made in the end.

2 Method

2.1 Participants

Volunteers were enlisted from the university. Twelve healthy students (7 males and 5 females with the average age of 22.5) who were with high interest to the study were selected. All of them were fully aware of what will be done in the experiment and signed the consent form before the experiment. They were rewarded for their participation after the experiment.

2.2 Apparatus

The experiment is carried out on a simulator with high fidelity. The simulator consists of two parts, the outside view and the cockpit. The outside view is simulated and projected on a cylindrical screen which has a diameter of about 8 meters. In the cockpit, the arrangements are referred to Boeing 777-200ER. There are control instruments and display instruments in the cockpit. The control instruments include the yoke, throttle, rudder pedal, flaps, landing gear, CDU and MCP. The display instruments include PFD, ND, EICAS, etc. The simulator can also record parameters such as the position of the aircraft, the speed, inputs of the control instruments and so on during the flight with the sample rate of 30 Hz[11].

A desk mounted eye tracker (Smart Eye Pro[12]) is used in the simulated flight experiment to measure the eye activities of the participant. This set of instruments can measure the participant's pupil diameter, blink, saccade and fixation with the frame rate of 60Hz.

Heart rate, respiration rate, respiration depth and body acceleration are measured by using the Zephry Bioharness system[13]. The data are recorded with the sample rate of 1Hz.

2.3 Task

In the experiment, each participant is asked to fly a complete flight task for 5 times. The flight task consists of take-off phase, cruise phase, approach and landing phase. The aircraft would take off from KSJC 30R. After passing 5 way points (i.e. SUNNE, WETOR, DUMBA, UNUWY and CEPIN, respectively) the aircraft would be landed at KSFO 28R. The flight environment is simulated as in summer, at noon, sunny and no wind.

Notice that in this experiment, the task difficulties in the three flight phases are different. Specifically, in the take-off phase, the participants only have to control the throttle and the elevator to let the aircraft climb. In the cruise phase, autopilot will be engaged and it will take the control of the aircraft and guide the aircraft to follow the preset route. In this phase, there will be no need for the participants to interfere with control of the aircraft by any manual control. In the approach and landing phase, we ask the participants to disengage the autopilot and perform a non-precision visual approach. The participants have to control the elevator, the aileron, the throttle, the flaps, the speed brakes and the landing gear to achieve this task. It requires not only many controls but also a lot of perceptions and decision making. Thus it's obvious that the workload imposed on the participant in these three flight phases was different.

2.4 Procedure

Before the experiment started, it took us three weeks to train the participants. Firstly, two tutorials were given to the participants to introduce the purpose of the experiment, some fundamental knowledge about the aircraft control and the measurement device used in the experiment. In the rest of the time, we scheduled a set of flight trainings. Each participant was trained in the simulator three times a week to execute the flight task which is identical to the task that has been described above. In each time, the training lasted an hour. And during the training, all the measurement devices are calibrated and deployed. All the objective data are collected during the training for the purpose of finding any potential problems which could occur during the experiment. Till the end of the training section, each participant achieved at least 9 hours of flight training.

After the training section, the experiment started. During the experiment, participants came to the lab according to the schedule. After arriving, they would first have 10 minutes to relax. Then the experimenter tested the devices and the participant entered the simulator. Next, the experimenter started recording and the participant began to perform the task. During the task the experimenter would check for the data to see if the devices were working properly. And if there were some problem with the devices, the experimenter would abort the experiment. After the aircraft had landed the experimenter stopped the recording and saved the data. The participant stayed in the simulator to have a 10 minutes break until the next trial started.

2.5 Analysis

Preliminary Processing. As mentioned before, the physiological data are recorded by several devices with different sample rate. For this problem, a two-step strategy is used to synchronize the measured data. Firstly, all measurement devices are connected to a local network. Before the experiment, the system time of each device is synchronized to a time server which is in the same local network. Secondly, after the experiment had completed, all measured data were gathered. Some of them were resampled (eye data), and some of them are interpolated (physiological data such as heart rate and respiration). Then these data were aligned and integrated according to the timestamps at each sample point by using a C++ program we developed. Notice that the timestamps of different data cannot be aligned strictly because they were actually not sampled at the same time. Therefore, a time interval of 20 milliseconds was set and if the timestamps difference is within that interval the data were assumed to be sampled at the same time. The final integrated data have the frame rate of about 30Hz.

Some common problems for psychophysiological measures are noise and lost detection. Noise can be introduced by many reasons such as the environmental interference, while lost detection of the data is often due to the poor contact of the sensor to the body. In the present study, interpolation is executed at where psychophysiological data are lost. Then a ten seconds Hanning window is used to filter and smooth the psychophysiological data[14]. For any psychophysiological feature, if over 10% of the data are lost in a trial, then this trial is discarded.

Because of the differences between the participants, absolute value of the physiological data were meaningless[15], all the data were transform to z-scores. The z-scores of the physiological data were used in the following statistical analysis.

Statistical Analysis. Several studies have shown that psychophysiological parameters are relevant to the task requirement. But due to the complex mechanism of the psychophysiological activities, the results were not always consistent with each other. In the present study, statistical analysis was made to determine the relationship between the physiological parameters and the task requirement.

It should be noted that in each flight phase, we extracted a segment of data for statistical analysis. For the take-off phase, the extracted segment covered the time interval from 60th second before the autopilot was engaged to the 30th second before the autopilot was engaged. For the cruise phase, the extracted segment covered the time interval from 180th second after the autopilot was engaged to 210th seconds after the autopilot was engaged. For the approach & landing phase, the extracted segment covered the last 30 seconds before the aircraft reached 2 nautical miles to the terminal airport. Average values of the physiological parameters in these segments were calculated. Then one-way ANOVA was carried out to test the sensitivity of each physiological parameter among the three segments.

3 Experimental Results

Sixty trials were completed in the experiment. Fifty-three data sets were used for analysis. For the other seven data sets, three of them were rejected because of the bad contact of the skin conductivity sensor and the heavy loss of detection, four data sets were rejected because of the malfunction of the eye tracker which stopped recording in the middle of the trial.

Table 1. ANOVA results

	F(2,156)	Sig.
Fixation duration	9.751	.000
Saccade frequency	.050	.952
Blink latency	3.689	.027
Blink duration	4.375	.014
Pupil diameter	18.383	.000
Heart rate	40.848	.000
Respiration rate	39.401	.000
Respiration depth	5.461	.005

From table 1, it could be indicated that fixation duration, pupil diameter, heart rate and respiration rate showed very significant differences ($p<0.001$) among take-off segment, cruise segment and approach & landing segment. For blink duration, blink latency and respiration depth, the differences among the three segments were also significant ($p<0.05$). But for saccade frequency, it didn't show significant difference among the three segments ($p>0.05$). Thus, saccade frequency cannot distinguish the difference of task difficulty and wouldn't be used in the following analysis.

Post hoc analysis using the Tamhane's T2 method (shown in table 2) indicated that pupil diameter and respiration rate significantly differed between each two of the three segments ($p<0.05$). Fixation duration didn't show significant difference between cruise segment and the approach segment ($p>0.05$). Heart rate didn't show significant difference between the climb segment and cruise segment ($p>0.05$). Respiration depth didn't show significant difference between the climb segment and approach segment ($p>0.05$). Blink duration only showed significant difference between the cruise segment and the approach segment ($p<0.05$). Blink latency didn't show significant difference between any two of the segments.

Table 2. Multiple Comparisons

Dependent Variable	(I) phase	(J) phase	Mean Difference (I-J)
fixation_duration	climb	cruise	.38506*
		approach	.50159*
	cruise	climb	-.38506*
		approach	.11653
	approach	climb	-.50159*
		cruise	-.11653
blink_latency	climb	cruise	.28874
		approach	.35108
	cruise	climb	-.28874
		approach	.06234
	approach	climb	-.35108
		cruise	-.06234
blink_duration	climb	cruise	-.12001
		approach	.17646
	cruise	climb	.12001
		approach	.29647*
	approach	climb	-.17646
		cruise	-.29647*
pupil_diameter	climb	cruise	.28085*
		approach	.72693*
	cruise	climb	-.28085*
		approach	.44608*
	approach	climb	-.72693*
		cruise	-.44608*
heart_rate	climb	cruise	-.12616
		approach	-.77144*
	cruise	climb	.12616
		approach	-.64528*
	approach	climb	.77144*
		cruise	.64528*

Table 2. (*continued*)

respiration_rate	climb	cruise	.30593*
		approach	-.65843*
	cruise	climb	-.30593*
		approach	-.96436*
	approach	climb	.65843*
		cruise	.96436*
respiration_depth	climb	cruise	-.29870*
		approach	.00371
	cruise	climb	.29870*
		approach	.30241*
	approach	climb	-.00371
		cruise	-.30241*

*. The mean difference is significant at the 0.05 level.

The statistical results of psychophysiological features showed that except saccade frequency, all features can distinguish the task difficulty to some extent. But the sensitivity of the physiological features was different.

4 Discussion

Physiological parameters including fixation duration, pupil diameter, blink duration, blink latency, saccade frequency, heart rate, respiration rate and respiration depth were measured during the flight task. The ANOVA results in table 1 indicated that saccade frequency didn't show significant differences among the three segments. The original assumption that saccade frequency would change along with task difficulty was found to be inadequate. During the approach and landing phase, the participant had to frequently get information from the attitude indicator, speed indicator, altitude indicator, navigation and out of the window, thus the saccade frequency was high. While during the cruise phase, when the participant didn't had to frequently get information and most of them explained to feel a little bit bored, it was found that the participant's eyes would still have many subconscious movements. Thus it can be inferred that saccade frequency was not sensitive to the task difficulty in this experiment. Meanwhile, considering the pair-comparison results in table 2, it indicated that the sensitivities of the participants' blink parameters (blink duration and blink latency) were not so satisfactory. By investigating the videos and the raw blink data recorded by the eye tracker, it was found that many of the blinks are actually "transition blinks" which accompanied with large range of eye movements. The mechanism of transition blinks is different from the ordinary blinks and they would affect the assessment of the blink-related features.

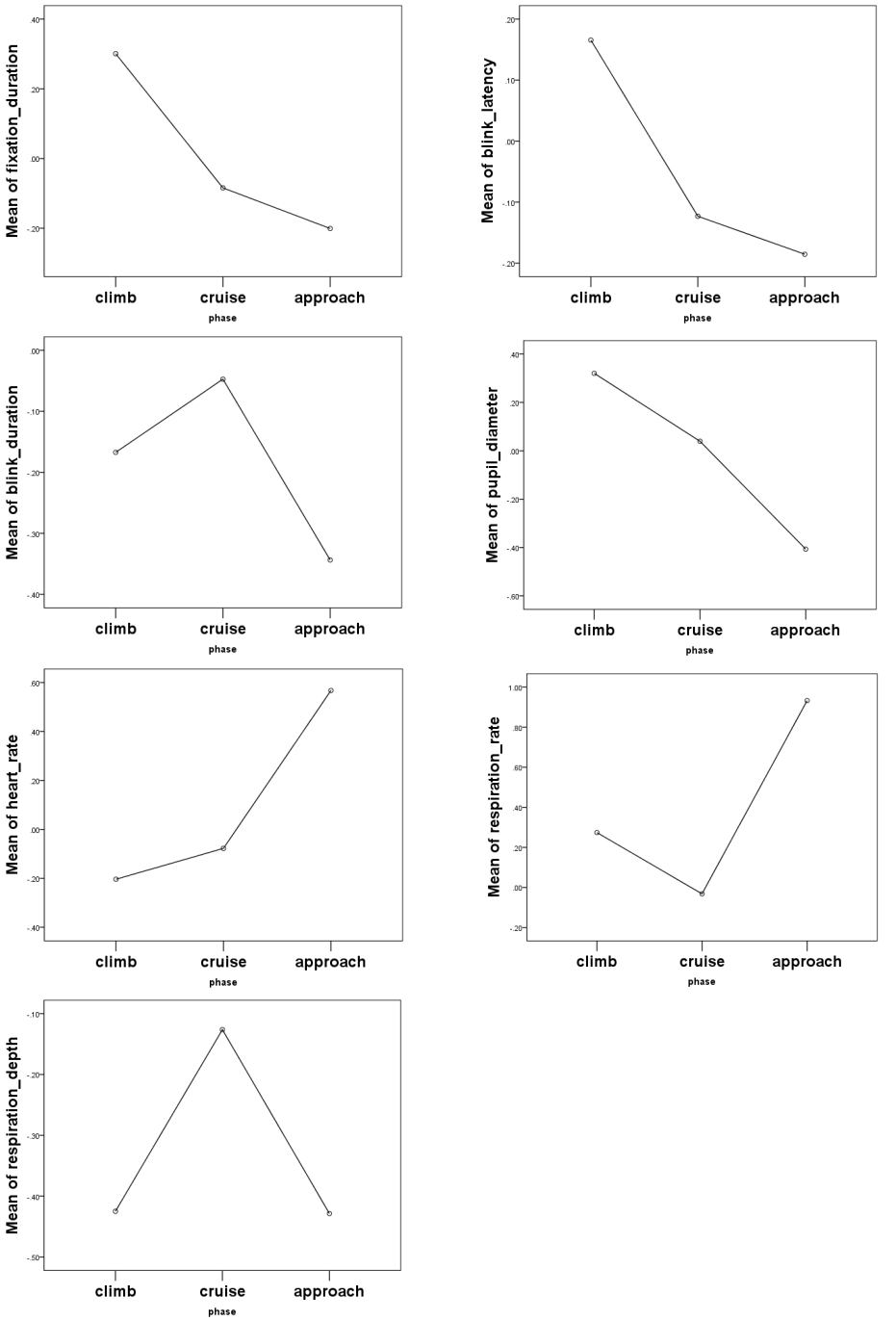


Fig. 1. Mean of the physiological parameters in different flight phases

Though the rest of the physiological features showed significant differences among the segments, their diagnosticity were still remained to be discussed. Fig. 1 showed the average value of the physiological features of all the 53 trials in different task segments. It indicated that some features changed in a "V" or inversed "V" style, such as respiration rate, respiration depth and blink duration. While for other features, some kept increasing such as heart rate; some kept the tendency to decrease such as fixation duration, pupil diameter and blink latency. It's easy to infer that the features changed in "V" or inversed "V" styles are related to task difficulty because the task difficulty in the three segments also had the "V" style. But heart rate, fixation duration, pupil diameter and blink latency seemed to be related to other factors. Since they had the accumulative effect, it could be explained that they were related to fatigue. Moreover, the changes of the fixation duration could also relate to the requirement of the task. For during the climb phase, participants had to pay most of their attention on the attitude indicator, thus the fixation duration were long. During the cruise phase, participants had certain amount of eye movements due to curiosity or boredom, thus they didn't fix on something for very long. While during the approach & landing phase, because of the high time pressure and the large amount of information requirements, the participants had to frequently glance at the indicators and the outside view, this made the fixation duration become very short.

5 Conclusions

In this paper, several physiological parameters were measured in a simulated flight task. According to the experimental results, it indicated that fixation duration, blink latency, blink duration, pupil diameter, heart rate, respiration rate, respiration depth were sensitive to the differences of task difficulty. Where pupil diameter and respiration rate were the most sensitive parameters, they could distinguish the differences of task difficulty between any two flight phases. On the other hand, the diagnosticity of the parameters was found to be different. Some parameters could be direct indicators of task difficulty, such as respiration rate, respiration depth and blink duration; others parameters in this experiment might be affected by not only the task difficulty but also other factors such as fatigue and so on. These results could provide some reference to further study in the assessment of the pilot's workload.

Acknowledgement. This research work was supported by National Basic Research Program of China-(973 Program No. 2010CB734103) and research program of Shanghai Jiao Tong University for innovation of post graduates (985 Program) with No. Z-413-006 under Grant TS0220741301.

References

1. Zolghadri, A.: Early warning and prediction of flight parameter abnormalities for improved system safety assessment. Reliability Engineering & System Safety 76(1), 19–27 (2002)

2. Hollands, J.G., Wickens, C.D.: Engineering psychology and human performance. Prentice Hall, New Jersey (1999)
3. Cain, B.: A review of the mental workload literature, DTIC Document (2007)
4. Harris, D.: Human Performance on the Flight Deck. Ashgate Publishing (2011)
5. Wang, Z., Fu, S.: A Layered Multi-dimensional Description of Pilot's Workload Based on Objective Measures. In: Harris, D. (ed.) EPCE 2013, Part II. LNCS, vol. 8020, pp. 203–211. Springer, Heidelberg (2013)
6. Stanton, N.A., et al.: Human factors methods: a practical guide for engineering and design. Ashgate Publishing (2012)
7. Farmer, E., Brownson, A.: Review of workload measurement, analysis and interpretation methods. European Organisation for the Safety of Air Navigation 33 (2003)
8. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems 6(2), 156–166 (2005)
9. Wilson, G.F.: An analysis of mental workload in pilots during flight using multiple psychophysiological measures. International Journal of Aviation Psychology 12(1), 3–18 (2002)
10. Veltman, J.A., Gaillard, A.W.K.: Physiological workload reactions to increasing levels of task difficulty. Ergonomics 41(5), 656–669 (1998)
11. Basler, M., Spott, M., Buchanan, S.: The FlightGear Manual (2010)
12. AB, S.E.: Smart-Eye Pro 5.6 User manual. Sweden-, Gothenburg, Smart Eye AB, Sweden Smart Eye AB (2009)
13. Technology, Z.: BioHarness™ User Guide (2010)
14. 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. International Journal of Industrial Ergonomics 35(11), 991–1009 (2005)
15. Corwin, W.H., et al.: Assessment of crew workload measurement methods, techniques, and procedures: Process, methods and results. Report no. WRDC-TR-89-7006 (1989)

A Theoretical Model of Mental Workload in Pilots Based on Multiple Experimental Measurements

Zongmin Wei, Damin Zhuang, Xiaoru Wanyan, Huan Zhang, and Chen Liu

Department of Human-machine Environment Engineering, School of Aeronautic Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing, China
weizongmin8411@163.com

Abstract. The present study attempted to establish an effective discrimination and prediction model that can be applied to evaluate mental workload changes in human-machine interaction processes on aircraft flight deck. By adopting a combined measure based on primary task measurement, subjective measurement and physiological measurement, this study developed both experimental measurement and theoretical modeling of mental workload under flight simulation task conditions. The experimental results showed that, as the mental workload increased, the peak amplitude of Mismatch negativity (MMN) was significantly increased, SDNN (the standard deviation of R-R intervals) was significantly decreased, the number of eye blink was decreased significantly. Finally, a comprehensive mental workload discrimination and prediction model for the aircraft flight deck display interface was constructed by the Bayesian Fisher discrimination and classification method. The model's accuracy was checked by original validation method. When comparing the prediction and discrimination results of this comprehensive model with that of single indices, the former showed much higher accuracy.

Keywords: Mental workload, Human-machine interaction, MMN, SDNN, Eye blink.

1 Introduction

Mental Workload has always been deemed as an important factor that influences pilots' performances, because under excessive mental workload, the pilots may exhibit delayed response to some of the incoming information; while under inadequate mental workload, the pilots may lack vigilant awareness or even miss some abnormal flight information. Aircraft flight deck display interface, as an important human-computer interaction of providing flight information to pilots, its design should provide appropriate mental workload, otherwise may seriously affect flight safety.

Accordingly, there are multiple methods for the measurement and evaluation of mental workload on the aircraft flight deck display interface, among which NASA_TLX scale subjective evaluation method is mostly widely used. NASA_TLX scale has the advantage of comprehensive evaluation, and it takes six factors into consideration, i.e. mental demand, physical demand, temporal demand, effort,

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 104–113, 2014.

performance and frustration level [1]. However, this method could only be conducted after the flight tests, which brings difficulties to the measurement and evaluation of aircraft cockpit human-machine interaction system design in earlier time, because once problems are detected, the system needs to be redesigned and the flight test and NASA_TLX scale subjective evaluation needs to be carried out once again, which puts great pressure on human and material resources. In addition, the evaluation results of NASA_TLX scale may be confusing due to the subjective differences among individuals [2].

Physiological evaluation method is another important measurement and evaluation method of mental workload on the aircraft flight deck display interface, which could provide real-time and objective physiological change of operators to measure mental workload. There are generally three categories of physiological evaluation methods: electroencephalogram (EEG), electrooculogram (EOG), and electrocardiogram (ECG) measurements. Former studies about flight simulation tests demonstrated that MMN index of EEG measurement [3], SDNN index of ECG measurement [4], as well as eye blink number of EOG measurement [5] could effectively reflect the mental workload level of flight, respectively.

However, current studies show that any single index or single method for the measurement and evaluation of mental workload has its own advantages and limitations [4]. Each index or method may provide useful information reflecting mental workload, but none could comprehensively reflect the mental workload under different task conditions. The multi-dimensional nature of mental workload makes each index sensitive to only one or several dimensions of mental workload, instead of all dimensions. That is to say, it is impossible for a single mental workload index attained from a single evaluation method to be appropriate for all different task conditions. Therefore, joint measurement and evaluation of a variety of indices of mental workload is seen as a more feasible approach and one of the future trends of mental workload evaluation.

The present study attempts to evaluate flight mental workload by synthetically combining above three categories of physiological measurements based on designed flight test tasks. It provides real-time and objective evaluation of the mental workload which the aircraft flight deck display interface puts on the pilots under different flight task conditions, and discriminates the mental workload level based the Bayesian Fisher discrimination analysis method, thereby guides the aircraft cockpit human-computer interaction systems mental workload task design. It is expected that the present study can be applied to the early design stage of aircraft flight deck display interface, and the designers can optimize the aircraft flight deck display interface by adjusting the mental workload task design, based on the discrimination and predication of mental workload.

2 Methods

2.1 Subjects

14 male flying cadets (range of age: 22-28 years old; mean of age: 24.6 years old) from Beijing University of Aeronautics and Astronautics participated in the present study. All subjects were right-handed with normal or corrected-to-normal hearing and

vision. Each subject was well trained to be good at simulated flight operations. After a complete description of the study, informed written consent was signed by each subject before the experiment.

2.2 Experimental Task

The experiment task was to monitor indicators based on a flight simulator. Subjects were required to accomplish the whole dynamic process of flight simulation, including take-off, climb, cruise, approach and landing. During the flight simulation process, they were asked to monitor the status of flight indicators presented on the Head up displays of simulation model, and recover the information state when abnormal information was detected, by pressing certain keys as quickly and accurately as possible. The simulation model was designed with reference to the typical information layout of Head up displays, and could display several kinds of flight indicators, including pitching angle, air speed, altitude, heading angle, rolling angle, and fuel status, as shown in table 1. Subjects' mental workloads were manipulated by adjusting the quantity of flight indicators and information refresh frequency: high mental workload was set as 6 flight indicators, 2s duration of abnormal information, and random inter-stimulus interval between abnormal information; low mental workload was set as 3 flight indicators, 1.5s duration of abnormal information, and random inter-stimulus interval between abnormal information.

Table 1. The scope setup for abnormal flight indicators

No.	Flight indicators	Abnormal scope
1	Pitching angle	Exceed 10'
2	Air speed	Exceed 400nautical mile/h
3	Altitude	Exceed 10000 feet
4	Heading angel	Exceed 50'
5	Rolling angel	Exceed 20'
6	Fuel status	Abnormal

2.3 Experimental Procedure

Within-subject factorial design was implemented in the experiment. The mental workload was divided into three levels, i.e. high, low and baseline. All the 14 subjects participated in the flight simulation experiments under the three levels of the mental workloads, respectively. The order of the high and low experiment tasks within the sessions was counterbalanced across participants to minimize the learning effect. In order to record the EEG, EOG and ECG data, all the participants were asked to wear EEG electrode cap, EOG electrodes and ECG electrodes throughout the experiment. After each session, each subject was instructed to take a 15-minnuts rest, meanwhile completed subjective evaluation of NASA_TLX (Task Load Index).

2.4 Experiment Data Recording and Analysis

Three types of indices were attained, including the subjects' performance evaluations (accuracy of the primary task and reaction time), subjective evaluations (the score of NASA_TLX), and three different physiological evaluations (EEG, ECG and EOG). By analyzing the changes of these different indices under different mental workload conditions, this study explored the sensitivity and diagnosability of these different indices to pilot mental workloads and provided a foundation for the design of display interface mental workload task.

Performance data recording and analysis. By computer programming, the system automatically recorded the two indicators of performance evaluation, including operation accurate rate and reaction time (the time interval between the detection of the abnormal information and correct responding).

Physiological data recording and analysis.FX-7402 12-channel automatic analysis of ECG machine was adopted to synchronously record the ECG signals. The ECG data recorded included the heart rates of subjects measured every 5 min, time series during R-R period, ECG within this period and the electrode placement arranged as per lead II. The heart rate value range was 20~300bpm, the heart rate detection accuracy was ±2bpm, the sampling frequency was 0.05-150Hz, and the waveform recording speed was 25mm/s. Relevant studies showed that, standard deviation of normal R-R intervals (denoted "SDNN"), one of the time-domain related indexes of HRV, could effectively reflect the sensitivity levels of mental workload[4]. Therefore, the present study analyzed the index of SDNN.

EOG signals were recorded from electrodes placed above and below the left eye. Relevant studies showed that eye blink number was closely related to mental workload level. Therefore, eye blinks data were analyzed in present study.

EEG signals were recorded from FZ electrode site using Neuroscan Nuamps Amplifier. Electrode placed at the forehead was grounding. Electrode impedances were maintained below 5KΩ. The recording band-pass was 0.05~100Hz and the sample rate was 1000 Hz. After correcting the eye movements, the epoch was set as 1300 ms, including a 200ms pre-stimulus baseline. Any epoch containing residual artifact voltages exceeding ±150μv was rejected. Relevant studies showed that the peak amplitudes of MMN of EEG measurement could effectively reflect the sensitivity of mental workload [3]. Therefore, the present study adopted the peak amplitudes of MMN of EEG。

Subjective data recording and analysis. NASA_TLX score was used for subjective analysis. For the convenience for the subjects to accurately and effectively complete subjective evaluation, in the present study, the NASA_TLX scale was presented in numerical value, i.e. score from 0 to 100, with 0 representing no effort and 100 representing maximum effort. First, a score (from 0 to 100) was obtained on each dimension according to the subjects' subjective feelings on the flight related mental workload. Then, a paired comparison task was performed for all pairs of the six dimensions, which required the subjects to choose which dimension had a greater

relevance to the overall mental workload. After that, each of the six dimensions was given a specific weight according to the number of times each dimension was chosen in pared comparison and the six dimensions were sorted. The final mental workload score was got by multiplying each individual dimension scale score by its respective weight and dividing the total score of all dimensions by 15 (the total number of paired comparisons). Repetitive measure analysis of variance (ANOVA) was employed for the analysis of the above data by using SPSS 17.0 statistical package.

3 Results

3.1 Flight Task Performance Measurement

At two different mental workload levels (high and low levels), the accuracy rate and reaction time of subjects responding to abnormal flight indicators were shown in Table 2. Two (high and low) × two (accurate rate and reaction time) repeated measure analysis of variance (ANOVA) showed a significant ($P<0.001$) main effect of mental workload. As the mental workload increased, the performance level decreased significantly ($P<0.001$), which was specifically demonstrated by the successive decrease ($P<0.001$) of the accuracy rate of participates as well as the successive increase ($P<0.001$) of the reaction time at high and low mental workload levels, respectively.

Table 2. Flight task performances under the high and low mental workloads

Mental workload	High	Low
Accuracy rate /%	74.14±5.67	97.88±1.75
Reaction time /ms	862.47±52.67	809.18±68.52

3.2 Subjective Measurement

Results of NASA_TLX-based subjective evaluation were shown in Table 3. Result of the single-factor repeated measure ANOVA suggested a remarkable ($P<0.001$) main effect of mental workload. With the increase of mental workload, the subjective evaluation scores of NASA_TLX gradually increased ($P<0.001$).

Table 3. Subjective measurement under the high and low mental workloads

Mental workload	High	Low
NASA_TLX	65.39±5.27	57.10±4.78

3.3 Physiological Measurement

For the peak amplitudes of MMN at Fz, the main effect of mental workload was significant ($p=0.008$). As the mental workload increased, the peak amplitudes of MMN decreased significantly. The result of a further paired comparison suggested that, the

peak amplitudes of MMN at high mental workload was significantly higher than that at low mental workload (P=0.035).

For the SDNN index, the main effect of mental workload was significant ($p<0.001$). As the mental workload increased, the value of SDNN decreased significantly. The result of a further paired comparison suggested that, the value of SDNN at high mental workload was significantly lower than that at low mental workload (P=0.013).

For the eye blinks index, the main effect of mental workload was significant ($p=0.002$). As the mental workload increased, the number of eye blinks decreased significantly. The result of a further paired comparison suggested that, the number of eye blinks at high mental workload was significantly decreased than that at low mental workload (P=0.003).

4 Modeling

4.1 Modeling Method

Based on the analysis results of experimental measurements, a comprehensive mental workload discrimination and prediction model for the aircraft flight deck display interface was constructed by the Bayesian Fisher discrimination and classification method. To ensure the comprehensiveness of the discrimination, the general discrimination analysis method (all-factor analysis method) was employed in the present study, i.e., the discrimination model included the peak amplitude of MMN, the value of SDNN, and eye blink number. The discrimination equations of the model included two discrimination functions, respectively representing two different mental workload levels. Substitute various index values obtained under the aircraft cockpit display interface successively into the two discrimination functions to calculate three scores, and the largest score represents the corresponding mental workload level.

4.2 Validity Check of the Model

The original validation method was used to check the predication and discrimination accuracy of the constructed Bayesian Fisher discrimination function. It substituted the 26 groups of subject sample data measured back into the constructed discrimination function to evaluate the accuracy level of predication and discrimination, and showed the check results in Table 4.

Table 4. Results of discrimination and predication accuracy rate

Method	**Predicted Mental Workload Accuracy Rate (%)**			
	Mental Workload	Low	High	Total
Original	Low	69.23	30.77	100.0
	High	15.38	84.62	100.0

4.3 Establishment of the Model and Instructions

It could be known from the comparative results of Table 4 that, when employing the general discrimination analysis method, the average discrimination and prediction accuracies of original check method was 76.92%. Specifically, the discrimination and prediction accuracies for low workload and high workload were 69.23% and 84.62%, respectively. The discriminate functions are as follows:

$$y_1=0.406x_1 - 0.075\ x_2 + 0.037\ x_3 - 12.666 \quad (1)$$

$$y_2=0.307\ x_1 - 0.222\ x_2 + 0.024\ x_3 - 8.149 \quad (2)$$

In the equations, y_1, y_2 represent the discriminate function value of the low and high levels of mental workloads, respectively. And x_1 value represents the SDNN value, x_2 value represents the peak amplitude of MMN, x_3 value represents eye blink numbers. According to the values of x_1, x_2, and x_3, the values of y_1, y_2 were calculated and compared. If the y_1 value is bigger, subjects are considered at a low level of mental workload. If the y_2 value is bigger, subjects are considered at a high level of mental workload.

5 Discussion

The major findings of the present study can be summarized as follows: as the mental workload increased, the three categories of indices represented different changes. For the performance evaluation indices, the detection accuracy of flight operation abnormal information was significantly decreased, and the response time was extended remarkably. For the subjective evaluation indices, the score of NASA_TLX was significantly increased. For the EEG evaluation indices, the peak amplitude of MMN was significantly increased; for the ECG evaluation indices, SDNN was sensitive to the mental workload change, which was significantly decreased as the mental workload increased; for the EOG evaluation indices, the number of eye blink was decreased significantly.

5.1 Sensitivity of Performance and Subjective of Indices to Mental Workload Change

The behavioral results indicated that the task performances of the subjects had been clearly distinguished between the high and low mental workloads conditions. Under the high mental workload condition, the accuracy rate of detecting abnormal information declined and the reaction time delayed. The outcomes implied that changing mental workloads during flight simulation condition affected subjects' operation performance significantly, and the results were consistent with the prior studies in other fields [6,7].

In the present study, the score of NASA_TLX increased significantly as the task got more difficult, which was consistent with former study results about flight tasks [5]. The present result also demonstrated that the setting of mental workload levels in this experiment for different flight task difficulties showed significant disparity from the participants' subjective perspective, which is accorded with the expectations.

5.2 Sensitivity of Integrated Evaluation Model to Mental Workload Change

The subjects' physiological indicators under different mental workload were chosen to construct the discrimination and prediction models of mental workload. Single physiological indicator evaluation model, two physiological indicators integrated evaluation model, and three physiological indicators integrated evaluation model, were detected by the Bayesian Fisher model to test their discrimination accuracy, respectively. The results are shown in Table 5.

Using original check method, by contrast, the integrated evaluation model based on three physiological indicators had the highest prediction accuracy (76.92%), followed by two physiological indicators integrated evaluation model by MS (MMN and SDNN) (73.08%) and SE (SDNN and Eye Blink) (69.23%), then was the two physiological indicators integrated evaluation model by EM (Eye Blink and MMN) (57.69%) and the single indicator evaluation model by Eye Blink index (57.69%), and finally the single indicator evaluation model by SDNN index (53.85%) and MMN index (53.85%).The overall comparison results of discrimination accuracy among different models showed that the discrimination accuracy of the model based on three physiological indicators is higher than the model based on two physiological indicators, which is higher than the model based on single physiological indicator. It demonstrated that multi-dimension physiological integrated evaluation model was generally more effective than single physiological indicator to discriminate the mental workload. In addition, among the single physiological indicator evaluation models, the Eye Blink indicator has the highest prediction accuracy for mental workload.

Table 5. Evaluation results of models based on single physiological indicator and multiple physiological indicators

Method	Evaluation indicators	Predicted Mental Workload Accuracy Rate (%)		
		Low	High	Total
Original	SDNN	46.15	61.54	53.85
	MMN	61.54	46.15	53.85
	Eye Blink	46.15	69.23	57.69
	SE (SDNN, Eye Blink)	61.54	76.92	69.23
	MS (SDNN, MMN)	76.92	69.23	73.08
	EM (Eye Blink, MMN)	61.54	53.85	57.69
	SME (SDNN, MMN, Eye Blink)	69.23	84.62	76.92

5.3 Comparison of Integrated Evaluation Model and NASA_TLX Score

The results of discrimination accuracy comparison between evaluation based on three physiological indicators integrated evaluation model and evaluation based on NASA_TLX score are shown in Table 6.

Using original check method, by contrast, under low mental workload, the evaluation based on three physiological indicators evaluation model had lower prediction accuracy than the evaluation based on NASA_TLX score; while under high mental workload, the evaluation based on three physiological indicators evaluation model had higher prediction accuracy than the evaluation based on NASA_TLX score; as for the average predication accuracy, the prediction accuracy of evaluation based on three physiological indicators evaluation model was the same as the evaluation based on NASA_TLX score. It demonstrated that it was feasible to substitute the three physiological indicators evaluation model for NASA_TLX score evaluation.

Table 6. Evaluation results of models based on three physiological indicators and NASA_TLX score

Method	Evaluation indicators	Predicted Accuracy Rate (%)		
		Low	High	Total
Original	SME (SDNN, MMN, Eye Blink)	69.23	84.62	76.92
	NASA_TLX	76.92	76.92	76.92

The overall comparison results of discrimination accuracy among different models showed that the discrimination accuracy of the model based on three physiological indicators is higher than the model based on two physiological indicators, which is higher than the model based on single physiological indicator. It demonstrated that multi-dimension physiological integrated evaluation model was generally more effective than single physiological indicator to discriminate the mental workload. In addition, among the single physiological indicator evaluation models, the Eye Blink indicator has the highest prediction accuracy for mental workload.

Therefore, the present study provides a method to extract the objective evaluation indicators sensitive to mental workload by experimental measurement, and then construct objective and real-time integrated discrimination and prediction model for mental workload change during flight process. It can help to determine and predict the mental workload task design for the aircraft cockpit human-machine interaction system design.

6 Conclusion

In conclusion, the discrimination and prediction results of the model proposed in this paper revealed a satisfactory consistency with the experimental measured results, and the model can accurately reflect the variation characteristics of the mental workload of the aircraft flight deck display interface, and provide a sound foundation for the ergonomic evaluation and optimization design of the aircraft flight deck display interface in the future.

Acknowledgments. This research is supported by National Basic Research Program of China (NBRPC, Program Grant No. 2010CB734104). The opinions and conclu-

sions expressed in this article are those of the authors and do not necessarily reflect the views of the NBRPC.

References

1. 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, Amsterdam (1988)
2. DiDomenico, A., Nussbaum, M.A.: Effects of different physical workload parameters on mental workload and performance. Int. J. Ind. Ergon. 41(3), 255–260 (2011)
3. Wei, Z., Wanyan, X., Zhuang, D.: Measurement and evaluation of mental workload for aircraft cockpit display interface. J. Beijing Univ. of Aero. and Astro. 40(1), 86–91 (2014)
4. Wei, Z., Zhuang, D., Wanyan, X., Liu, C., Zhang, H.: A model for discrimination and prediction of mental workload of aircraft cockpit display interface. Chi. J. Aero. (accepted)
5. 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)
6. Cohen, J., Polich, J.: On the number of trials needed for P300. Int. J. Psychophysiol. 25(3), 249–255 (1997)
7. Ullsperger, P., Freude, G., Erdmann, U.: Auditory probe sensitivity to mental workload changes–an event-related potential study. Int. J. Psychophysiol. 40(3), 201–209 (2001)

Long-Term Psychosocial Stress Attenuates Attention Resource of Post-Error

Yiran Yuan[1,2], Jianhui Wu[1], and Kan Zhang[1]

[1] Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
{yuanyr,wujh,zhangk}@psych.ac.cn

Abstract. This study examined both the behavioral performance and the brain mechanisms of post-error adjustments under long-term psychological stress by using ERP technique. Forty two participants who had been exposed to long-term exam preparation (versus 21 controls who were not exposed to such exam) performed a Go/NoGo task while electroencephalograms were recorded. We used Cohen's Perceived Stress Scale to assess their chronic stress level, and results suggested that participants in the exam group had higher levels of perceived stress. Although the behavioral performance of post-error trials had no difference between two groups, the exam group elicited significantly decreased P3 amplitude than the non-exam group in the post-error condition. Furthermore, the P3 amplitude in the post-error condition was negatively correlated with the perceived stress scores, suggesting that long-term psychosocial stress may lead to the decrease of the attention resource after committing an error.

Keywords: Long-term psychosocial stress, Post-error adjustments, Event-related potentials, Go/NoGo, P3, Attention resource.

1 Introduction

Human brain not only has the ability to detect errors but also to adjust behavioral performance after committing an error. One study has showed that stress can impair processes involved in error detection [1]. But how effect of stress on post-error behavior?

The common observed post-error adjustments reflected on change of reaction time and accuracy rate [2]. Usually we observed a prolonged reaction time responded after an error trial compared to a correct trial (so called post-error slowing) [3, 4]. The literatures have showed the mix results of accuracy rate in post-error. Some studies found that the accuracy increased after committing an error [5, 6], while some other studies found decreased or no affect [7, 8]. There are some theories to explain for post-error adjustments. One explanation is orienting account, which has received much attention nowadays. It refers to an orienting response elicited by error (infrequent events) and leads to prolonged reacting time in post-error trials, sometimes in combination with decreased accuracy [7, 9]. The previous ERP result that the P3 amplitude of error

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 114–122, 2014.

trials was positively correlated with the post-error reaction time also supported the orienting account [10].

It is well known that stress not only lead to an increased activity of the hypothalamic-pituitary-adrenocortical axis (HPA axis), but also impacts the cognitive function and emotion [11]. But how cognition and behavior are modulated after committing an error under the stressful situation? One behavioral study reported that baseline cortisol was independently positively associated with post-error slowing [12], suggesting that more stressed state before the task as indexed by the cortisol level could increase post-error slowing. However, little is known whether and how long-term stress affects the post-error adjustments.

The aim of this study is to examine both the behavioral performance and the neural dynamics of post-error adjustments under long-term psychological stress by using the ERP technique. We used a major, highly competitive Chinese National Postgraduate Entrance Exam (NPEE) as long-term psychological stressor. The participants performed a Go/NoGo task while EEG data was recorded. Perceived stress scale was obtained to assess the effect of long-term stressor exposure. According to the orienting account of post-error adjustments and the P3 component was associated with the attention resource allocation [13], we expect long-term psychosocial stress to decrease attention resource of post-error trials, reflected by the decreased behavioral performance and/or attenuated P3 amplitude of post-error (false-alarm of the NoGo trials) Go trials.

2 Material and Methods

2.1 Participant

Forty two young healthy participants who had exposed to a major, high-competitive Chinese National Postgraduate Entrance Exam (NPEE) for 6 months and 21 non-exam as controls were recruited for this study. Considering gender differences in the effects of stress [14, 15], only male participants were recruited in this study. All participants were assessed by the Chinese version of the Life Events Scale (LES) [16, 17] to exclude other major life stressors during the past month. This experiment was approved by the Ethics Committee of Human Experimentation at the Institute of Psychology, Chinese Academy of Sciences. All participants provided written informed consent and were compensated for their participation.

2.2 General Procedure

Between 11-25 days before the national NPEE, all qualified participants came to the laboratory, completed questionnaires and several psychological tests including Go/NoGo task while EEG data was collected (the other tests didn't report here).

2.3 Psychological Measurements

To assess chronic stress level, all qualified participants completed the Perceived Stress Scale (PSS 10-item version) [18], which were widely used as an index of the

perception of chronic stress [19, 20]. In addition, the participants completed the Chinese version of the Mini International Personality Item Pool (the Mini-IPIP) to measure the Big Five factors of personality (neuroticism, extraversion, conscientiousness, openness, and agreeableness) [21].

2.4 Go/NoGo Task

During each trial, participants were asked to respond as soon as possible to letter "O" (Go trial) by pressing the button with the index finger of one of their hands, while didn't respond to the letter "X" (NoGo trial). The buttons were counterbalanced for the left/right hand across the participants. The probability of Go trial and NoGo trial was 80%: 20%. The consecutive presentation of two NoGo trials was avoided. Before the experiment session, participants received a practice session of 20 stimuli. During the experiment session, all participants received two blocks each consisting of 240 stimuli with 1-2 min breaks between blocks. The stimuli were displayed for 500 ms with a random interstimulus interval of 1200–1500ms.

2.5 ERP Recordings

During the experiment session, electroencephalograms (EEG) were continuously recorded from 64 scalp sites using Ag/AgCl electrodes mounted in an elastic cap (Neuroscan Inc., USA). The ground electrode was placed on the forehead, with an on-line reference to the left mastoid and an off-line algebraic re-reference to the average of the left and right mastoids. The vertical (VEOG) and horizontal electrooculograms (HEOG) were recorded from two pairs of electrodes, one pair placed above and below the left eye, and another pair placed at 1 cm from the outer canthi of each eye. All interelectrode impedances were kept below 5 kΩ. The signals were amplified by a Neuroscan SynAmps^2 amplifier (Neuroscan Inc., USA) with a 0.05-100 Hz bandpass filter and digitized at 1000 Hz.

The EEG data were digitally filtered with a 30 Hz lowpass filter and epoched into periods of 1000 ms (including a 200 ms prestimulus baseline) that were time-locked to the onset of the presented digit. The EEG signal was corrected by removing ocular artifacts through a regression procedure implemented in the Neuroscan software [22]. Trials with various artifacts were rejected, with a criterion of ± 100 μV.

2.6 Data Analyses

We used independent sample t-tests to compare the differences of exam and non-exam group on PSS and the Mini-IPIP.

There were two conditions for analyses of both behavioral performance and ERP measures. Post-correct condition referred to the Go trial after the hit trial. Post-error condition referred to the Go trial after the false alarm trial. For the behavioral performance, the percentage of correct responses (correct rate) and reaction time of correct responses (RT) were calculated separately for post-error and post-correct condition. For the ERP data, the mean amplitude of P3 was measured in each condition during the time interval from 250 to 310 ms after stimulus onset at Pz site.

Repeated-measures ANOVAs with condition (post-correct vs. post-error) as the within-subjects factor and group (exam group vs. non-exam group) as the between-subjects factor were calculated separately for both behavioral performance and ERP data.

Participants who had less than 10 false alarm trials were excluded before data analyses. Finally, 35 participants in the exam group and 18 participants in the non-exam group remained when analyzing the behavioral performance. In addition, participants who had less than 10 accepted trials of ERP data were also excluded, so there were 27 participants in the exam group and 12 participants in the non-exam group remained when analyzing the ERP data.

Correlation analyses using Pearson's r were conducted between perceived stress and the behavioral performance and ERP data of post-error adjustments for all the participants.

The Greenhouse-Geisser correction was used to compensate for sphericity violations. Measures of effect size are reported using eta square (partial η^2). All p values $\leq$ .05 were considered statistically significant (two-tailed).

3 Results

3.1 Psychological Measurements of Long-Term Psychosocial Stress

The exam group and non-exam group were matched with respect to age (M ± SD: exam group 22.4 ± 1.0 years vs. non-exam group 22.7 ± 1.1 years). Perceived stress scores was significantly higher in the exam group than in the non-exam group (t = 2.197, df = 22, p = 0.039; see Fig 1). There was no significant difference between the exam group and the non-exam group on big five factors of personality ($p_s > 0.1$).

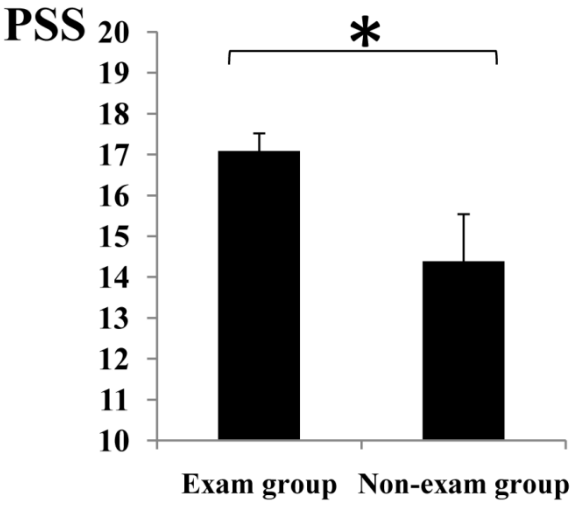


Fig. 1. Average Perceived stress scores (PSS) in the exam and non-exam groups. Error bars represent standard error of the mean. Notes: *: $p < 0.05$.

3.2 Effects of Long-Term Psychosocial Stress on Behavioral Performance

Analysis of correct rate revealed a significant main effect for condition (post-correct vs. post-error). In comparison with the post-correct condition, participants in both groups had significantly lower correct rates in the post-error condition ($F_{(1,51)}$ = 15.269, $p < 0.001$, partial $\eta^2 = 0.230$). Neither the main effect for group nor the interaction between the two factors was significant ($p_s > 0.1$). Analysis of RT showed that neither the main effects nor the interaction was significant ($p_s > 0.1$) (see Fig 2).

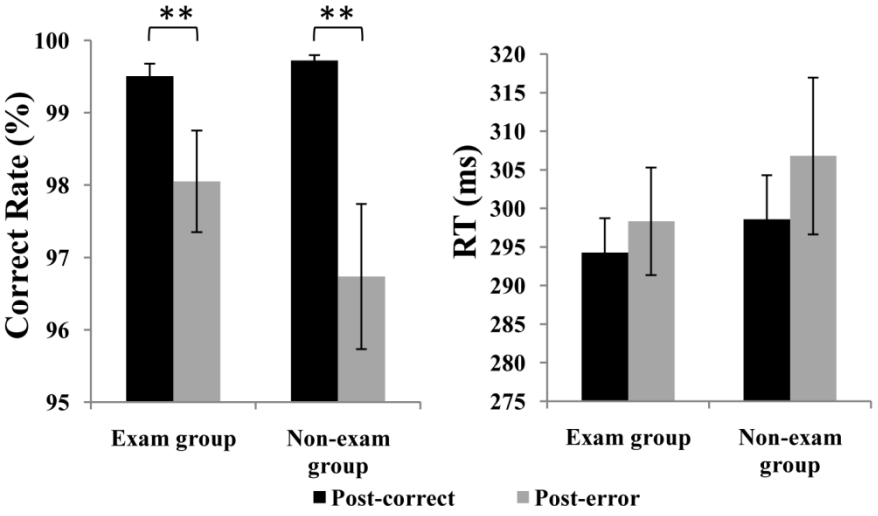


Fig. 2. Correct rate and RT for the post-correct and post-error conditions in the exam and non-exam groups. Error bars represent the standard error of the mean. Notes: **: $p < 0.01$.

3.3 Effects of Long-Term Psychosocial Stress on P3 Amplitude

The ANOVA revealed a significant main effect for condition and group (condition: $F_{(1,37)} = 7.307$, $p = 0.010$, partial $\eta^2 = 0.165$; group: $F_{(1,37)} = 4.424$, $p = 0.042$, partial $\eta^2 = 0.107$). The type × group interaction also reached significance ($F_{(1,37)} = 4.394$, $p = 0.043$, partial $\eta^2 = 0.106$). For further analysis of the interaction, there was no difference between the exam group and the non-exam group in the post-correct condition ($p > 0.1$). But in the post-error condition, the P3 amplitude of the exam group was significantly smaller than the non-exam group ($p = 0.015$) (see Fig 3).

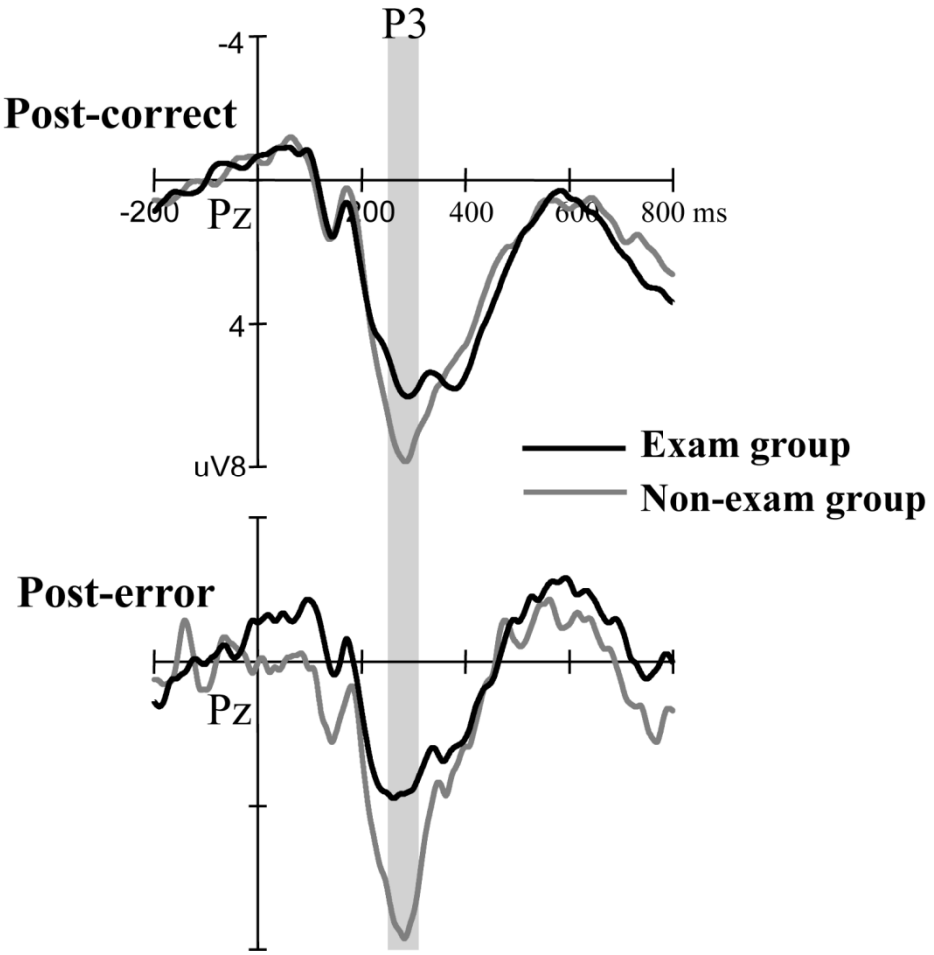


Fig. 3. Grand average ERPs elicited by performing post-correct and post-error conditions in the exam group and non-exam group at Pz electrode site. The gray areas highlight the time windows for P3 (250–310 ms) that was used for the statistical analysis.

3.4 Relationship between Perceived Stress and P3 Amplitude

For the whole participant sample, the perceived stress scores was negatively correlated with P3 amplitude in the post-error condition (r = -0.318, p = 0.048). There were no significant relationships between perceived stress and behavioral performance of post-error adjustments ($p_s > 0.1$).

4 Discussion

The present study investigated effects of long-term psychosocial stress on both the behavioral performance and the neural dynamics of post-error adjustments.

Psychological assessments confirmed that participants in the exam group were exposed to high levels of perceived stress. In comparison with the non-exam group, the exam group elicited a decrease in the P3 amplitude in the post-error condition when performing a Go/NoGo task, whereas behavioral performance remained no change between two groups. Furthermore, the P3 amplitude in the post-error condition was negatively correlated with the perceived stress scores.

Our behavioral results only showed that participants in both groups had significantly lower correct rates in the post-error condition of Go trials than in the post-correct condition. RT in the post-error condition was longer than in the post-correct condition, but it didn't reach significant level. These results were consistent with previous finding indicating that post-error slowing and correct rates of post-error were not always co-occur [2]. No change was found between two groups on behavioral performance, maybe the output of behavioral performance was not sensitive to stress.

Importantly, our ERP result showed significantly decreased P3 amplitude of the exam group compare with the non-exam group in the post-error condition. Literature has suggested that the P3 component was associated with the attention resource allocation [13]. Our results was consistent with the explanation of orienting account [9], suggesting that the exam group have less attention resource allocated to the post-error trials. Furthermore, the P3 amplitude in the post-error condition was negatively correlated with the perceived stress scores. It suggested that long-term psychosocial stress may lead to decreased attention resource in the post-error condition. The previous studies also have showed that stress can result in some cognitive consequences, such as attentional tunneling and impaired attention shifting [19, 23].

Although there was no group difference in behavioral performance of post-error behavior, the ERP result showed significantly difference between two groups. This might implicate that we cannot only depend on the final behavioral output when we examine the effects of stress on cognitive function. The event-related potentials (ERPs) technique is a widely used method to examine alterations in the dynamic time course, known as its high temporal resolution in millisecond. The current results suggest that chronic stress modulated the step of attention resource allocation for the post-error behavior. According to our knowledge, this is the first ERP evidence suggesting the stress may modulate the post-error behavior.

There are a few limitations in our study. First, small sample size were analyzed in this study partially due to the fact that we excluded some participants because they committed small number of errors. Second, we used only male undergraduate students as participants, which might limit the generalizability of our results. Third, we did not evaluate whether the two groups differed before they started preparing for the NPEE, so we cannot directly conclude that long-term psychosocial stress leads to a difference between the two groups on post-error adjustments. But we assessed the big five factors of personality on two groups and the results showed that no difference between two groups. A future study may add a test during a non-exam period to obtain a baseline to study the effects of the exam.

To summarize, our results provide electrophysiological evidence that long-term psychosocial stress may lead to the decrease of the attention resource after committing an error, which reflecting on significantly decreased P3 amplitude in the post-error condition of the exam group than the non-exam group.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (91124003, 81371203, 31100750 and 31100734).

Reference

1. Hsu, F.C., Garside, M.J., Massey, A.E., McAllister-Williams, R.H.: Effects of a single dose of cortisol on the neural correlates of episodic memory and error processing in healthy volunteers. Psychopharmacology (Berl) 167(4), 431–442 (2003)
2. Danielmeier, C., Ullsperger, M.: Post-error adjustments. Front. Psychol. 2, 233 (2011)
3. Danielmeier, C., Eichele, T., Forstmann, B.U., Tittgemeyer, M., Ullsperger, M.: Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. J. Neurosci. 31(5), 1780–1789 (2011)
4. Rabbitt, P.M.: Errors and error correction in choice-response tasks. J. Exp. Psychol. 71(2), 264–272 (1966)
5. Klein, T.A., Endrass, T., Kathmann, N., Neumann, J., von Cramon, D.Y., Ullsperger, M.: Neural correlates of error awareness. Neuroimage 34(4), 1774–1781 (2007)
6. Maier, M.E., Yeung, N., Steinhauser, M.: Error-related brain activity and adjustments of selective attention following errors. Neuroimage 56(4), 2339–2347 (2011)
7. Fiehler, K., Ullsperger, M., von Cramon, D.Y.: Electrophysiological correlates of error correction. Psychophysiology 42(1), 72–82 (2005)
8. Hajcak, G., Simons, R.F.: Oops!.. I did it again: an ERP and behavioral study of double-errors. Brain Cogn. 68(1), 15–21 (2008)
9. Notebaert, W., Houtman, F., Opstal, F.V., Gevers, W., Fias, W., Verguts, T.: Post-error slowing: an orienting account. Cognition 111(2), 275–279 (2009)
10. Nunez Castellar, E., Kuhn, S., Fias, W., Notebaert, W.: Outcome expectancy and not accuracy determines posterror slowing: ERP support. Cogn. Affect. Behav. Neurosci. 10(2), 270–278 (2010)
11. Lupien, S.J., Maheu, F., Tu, M., Fiocco, A., Schramek, T.E.: The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. Brain Cogn. 65(3), 209–237 (2007)
12. Tops, M., Boksem, M.A.: Cortisol involvement in mechanisms of behavioral inhibition. Psychophysiology 48(5), 723–732 (2011)
13. Isreal, J.B., Chesney, G.L., Wickens, C.D., Donchin, E.: P300 and Tracking Difficulty: Evidence for Multiple Resources in Dual–Task Performance. Psychophysiology 17(3), 259–273 (1980)
14. Backovic, D.V., Zivojinovic, J.I., Maksimovic, J., Maksimovic, M.: Gender differences in academic stress and burnout among medical students in final years of education. Psychiatr. Danub. 24(2), 175–181 (2012)
15. Weekes, N.Y., Lewis, R.S., Goto, S.G., Garrison-Jakel, J., Patel, F., Lupien, S.: The effect of an environmental stressor on gender differences on the awakening cortisol response. Psychoneuroendocrinology 33(6), 766–772 (2008)
16. Tennant, C., Andrews, G.: A scale to measure the stress of life events. Australian and New Zealand Journal of Psychiatry 10(1), 27–32 (1976)
17. Zhang, Y.L., Yang, D.S.: Life event scale. Chin. Mental Health 13, 101–103 (1999)
18. Cohen, S.: Perceived stress in a probability sample of the United States. In: Spacapan, S., Oskamp, S. (eds.) The Social Psychology of Health: Claremont Symposium on Applied Social Psychology, pp. 31–67. Sage Publications, Inc., Newbury Park (1988)

19. Liston, C., McEwen, B.S., Casey, B.J.: Psychosocial stress reversibly disrupts prefrontal processing and attentional control. Proc. Natl. Acad. Sci. U. S. A. 106(3), 912–917 (2009)
20. Tomiyama, A.J., Dallman, M.F., Epel, E.S.: Comfort food is comforting to those most stressed: evidence of the chronic stress response network in high stress women. Psychoneuroendocrinology 36(10), 1513–1519 (2011)
21. Donnellan, M.B., Oswald, F.L., Baird, B.M., Lucas, R.E.: The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. Psychological Assessment 18(2), 192 (2006)
22. Semlitsch, H.V., Anderer, P., Schuster, P., Presslich, O.: A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. Psychophysiology 23(6), 695–703 (1986)
23. Fredrickson, B.L., Branigan, C.: Positive emotions broaden the scope of attention and thought-action repertoires. Cogn. Emot. 19(3), 313–332 (2005)

Visual Perception

Reflection Overlay as a Potential Tool for Separating Real Images from Virtual Images in Photographs of Architecture

Marcin Brzezicki

Faculty of Architecture, Wroclaw University of Technology,
Prusa 53-55, 50-317 Wroclaw, Poland
marcin.brzezicki@pwr.wroc.pl

Abstract. The perception of transparency in architecture poses a major cognitive challenge due to the nearly faultless quality of large-scale light-permeable materials. Previous experiments demonstrated the significance of the virtual image in the perception of light-permeable materials. The separation of this image in photographs of architecture proved crucial in the research, giving the freedom to manipulate independently the components of the perceived scene. In the paper the author presents a new methodology of separating and manipulating real and virtual images, which is based on real-life photographs, not computer-generated scenes. The paper also proposes a step-by-step image processing algorithm which helps to better understand the cognitive processes of the human visual system, and presents sample results of this method.

Keywords: transparency perception, virtual image, digital image processing.

1 Introduction

The perception of transparency in architecture poses a major cognitive challenge due to the nearly faultless quality of large-scale light-transmitting panes. Smooth light-permeable panes simultaneously transmit and reflect the luminous fluxes that enter the panes from both sides, which, in turn, generates two separate luminous fluxes affecting the observer. Those luminous fluxes are formed at the stage of distal stimuli generation. When the proximal stimulus reaches the sensory receptor organs, these luminous fluxes become con-fused (deliberate hyphen) proportionally to their value. Since they stimulate the same area of the retina, they are further processed by the visual system as two superimposed images: the real one and the virtual one. The superposition of those two images constitutes a real challenge for the visual system. The amount of information reaching the retina is the same (ca. 100 billion bits/second), but the interpretative potential is multiplied.

The specular reflection off the smooth surface of the pane and the resulting virtual image are the main perceptual cues that the human visual system interprets to identify the optical transparency of the pane. The virtual image is also used to determine the orientation of the pane in the 3-dimensional space surrounding the observer, i.e. the

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 125–132, 2014.

so-called “cognitive map”. The cognitive process of attention carefully selects the information contained in the superimposed images, but, paradoxically, “establishing whether a given information is significant requires full processing” of said information [1, p. 115]. Conversely, research shows that if the virtual image is not present, the light-permeable pane seemingly vanishes and becomes invisible for the observer, who can accidentally walk into it and get injured. The American architect and theorist prof. Michael Bell, of the Columbia University’s Graduate School of Architecture, writes: “in an idealized realm, without reflection or surface complexity, glass has always strived to disappear” [2, p. 13]. Thus, it may be stated that the formation of the virtual image determines the user’s safety. The following research aims at furthering the knowledge on this issue by improving the understanding of the processes that occur.

2 Advancing the Research of Transparency Perception

The ubiquity of virtual images on smooth light-permeable panes requires one to determine specific conditions for study and experimentation. In previous research performed by the author, the virtual image, which was considered the main transparency cue, was verified through a negative control experiment. The experiment was based on a comparative estimation of the perception cues available to the human visual system before and after removing the virtual image from the observer’s field of view. This virtual image is removed when distal stimuli are formed – when fluxes interact with the panes, before the con-fused luminous fluxes affect the retina. The virtual image can be eliminated because when each ray of light is reflected (i.e. it bounces off), it changes the direction of its polarization. Such polarized light can be selectively blocked using appropriate optical filters.

The general concept of this research is based on two assumptions: *(i)* the influence of the virtual image on the perception of light-permeable panes could be more adequately measured by manipulating the virtual image, to check, how it affects the perception of transparency; *(ii)* the virtual/real image ratio is more credible and explicit in real-life photographs than on computer simulations[1]. The analysis of real-life photographs of architecture lacks the precision of the analytical method and requires careful preparation of the image data, but gives the researcher a full record of the actual light field – all the luminous fluxes affecting the real-life scene at a chosen moment, from a chosen angle.

3 Terms and Recent Studies Report

In his recent research (presented at the HCI 2013 conference) the author developed a photograph-based method called “pictorial image analysis” [3]. This method

[1] The amount of light which enters a building and which is reflected could be calculated analytically based on the optical parameters of the light-permeable materials, the plan of the building and the position of the viewer.

compares pairs of corresponding photographs of the same light-permeable pane, of which one is unmodified and contains the virtual image formed on the surface of the light-permeable pane (hereinafter called *file A*), and the second is without this image (hereinafter called *file B*). *File B* was photographed using a polarizing filter to block the reflected rays of light that make the virtual image "visible" on the surface of the pane. Afterwards, the pairs of corresponding images were digitally processed using image processing software.

4 Data Acquisition Procedure: *file A* and *file B*

The image data were acquired following a strict procedure, which ensured that the corresponding images in each pair had identical values of image parameters, such as exposure, color matrix, brightness curve etc.

4.1 Image Shooting

The pairs of corresponding files, i.e. *file A* and *file B,* were captured using Sony Alpha 100 DSLR camera with Sigma 10-20 mm lens. The camera was stabilized on a tripod to assure the same field of view; the shutter speed and the aperture values were manually set. A lens-mounted polarizing filter was used to block the reflection bouncing off the light-transmitting pane. The filter's rotation angle was set manually to achieve the best result, which was assessed visually on site. At least three images were recorded, each with different angles of filter rotation. The photos were taken as RAW images, i.e. unprocessed data directly from the CCD image sensor of the digital camera.

4.2 Image Post-processing

The RAW image data was processed using Image Data Converter SR ver. 1.1.00 by Sony, which is dedicated software for this camera model. The pairs of corresponding RAW images were converted to TIFF 24-bit true color with the same exposure values (EV) and other image parameters.

4.3 Case Study Buildings

Two case study buildings were selected for this stage of research: Thespian Housing and Office Building (Macków Pracownia Projektowa, 2010) and Wroclawski Park Wodny – a swimming pool complex (arch Horst Haag/I-Plan GmbH, 2008). The first building had been studied in the previous part of the research, published in HCII 2013 conference materials. The buildings were photographed from different vantage points and from different angles (see Fig. 1). The Thespian Housing and Office Building is prominent example of contemporary architecture, recognized worldwide, and it was a "nominee for the European Union Prize for Contemporary Architecture – the Mies van der Rohe Award" [3, p. 191].

a *b*

Fig. 1. Case study buildings: a) Thespian Housing and Office Building (Macków Pracownia Projektowa, 2010); b) Wroclawski Park Wodny (arch Horst Haag/I-Plan GmbH, 2008)

5 Image Subtraction

The visible image on the light-permeable pane is composed of two images: the real one, which is transmitted through the pane, and the virtual one, which is created by the rays of light reflected off the pane. These two component images: real and virtual, or, in other words, transmitted and specularly reflected, are transparently overlaid on the pane. This is the result of the con-fuse, or mix, of luminous fluxes which form the component images. If the luminance of the virtual image is higher than that of the real image, the real image is less visible than the virtual image. If the opposite happens – the virtual image is less visible. The original luminance values of the of the "weaker" images remain unaffected, i.e. darkening does not occur! The impression of attenuation of the "weaker" image results from the physiology of the human eye, which adjusts the size of the pupil in response to the amount of light present in the most intense luminous flux.

The proposed image processing algorithm bases on the analysis and comparison of color pixel values in *files A* and *B*. The observation is that the real image is less visible than the stronger virtual image, which – in terms of luminance and color values – means the brightening of the color resulting from the con-fuse. As a result of the con-fuse of luminous fluxes, the colors are mixed in an additive manner based on the RGB model values (0-255, 0-255, 0-255). Hence, it is possible to calculate the difference between *file A* with the virtual image, and *file B* without this image. This could be achieved by subtracting images from each other, where the RGB values are respectively subtracted, creating a new image, hereinafter called *subtraction image* (see Fig. 2).

The image calculation – i.e. the subtraction – was performed using ImageJ software. The obtained *subtraction image* was saved as 24-bit true color TIFF image. The calculation was carried out using the ImageJ post-production software for the analysis of medical image data [4], which was originally developed by the Research Service branch of the U.S. National Institutes of Health. This was followed by image processing operations (described below), which produced the *reflection overlay*.

Fig. 2. Phases of image subtraction: *file A* and *file B* and the resulting *subtraction image*, i.e. the separated reflection on a black background. The regions unaffected by the reflection are black, whereas the other regions contain different levels of luminance and color.

6 Creating the Reflection Overlay

The new concept presented in this paper, called *reflection overlay,* basically simulates a natural light-field by means of digital software. The method makes use of transparent layers, which are a common tool in digital image editing offered by popular and professional software. The layers are stacked on top of each other, and, depending on the order and the transparency settings, they determine the appearance of the final outcome.

The idea behind the *reflection overlay* method is to stack the elements of the image in the correct order. *File B* without the virtual image must always be placed at the bottom of the stack. A separate transparent *reflection overlay* layer with only the virtual image is placed directly over it. Because it is separated, it can be freely modified, duplicated or distorted without affecting the other layers. This makes it possible to observe the influence of the virtual image on the perception of light-permeable materials in architecture.

The next section presents two approaches to creating transparent *reflection overlay*, which are based on digital image post-processing, and which produce slightly different results.

6.1 Using the Alpha Transparency Channel

Additional image data can be stored in the so-called alpha channel (with a value between 0 and 1) and denote the transparency of the pixels: 0 meaning full transparency, and 1 meaning full opacity. In the *subtraction image*, the regions unaffected by the reflection are black, whereas the regions affected by the reflection contain different levels of luminance and color. A simple image processing operation: Layer → Transparency → Color To Alpha – converts this *subtraction image* into the transparent *reflection overlay*. This is achieved by assigning a channel value to each pixel based on their corresponding luminance values, e.g. a black pixel which becomes fully transparent has a channel value of 0 assigned to it; a white pixel which becomes totally opaque has a channel value of 1 assigned to it; intermediate values of alpha are assigned to other pixels depending on their luminance values (see Fig. 3).

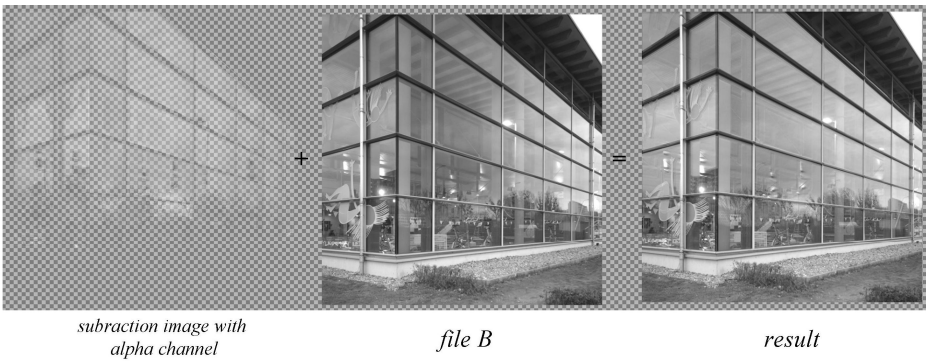


Fig. 3. Conversion of the *subtraction image* to the transparent *reflection overlay* using Layer → Transparency → Color To Alpha. The alpha channel value of 0 is assigned to every black pixel, while the values between 0 and 1 are assigned to other pixels, depending on their luminance.

6.2 Using the "Addition" Layer Mode

Similar results can be achieved by simply manipulating layer properties. This method requires the same order of layers as in the first approach, but the alpha channel transparency is not assigned to the top layer with the *subtraction image*. The transparency of the *reflection overlay* is achieved by setting the Layer Mode to "addition".

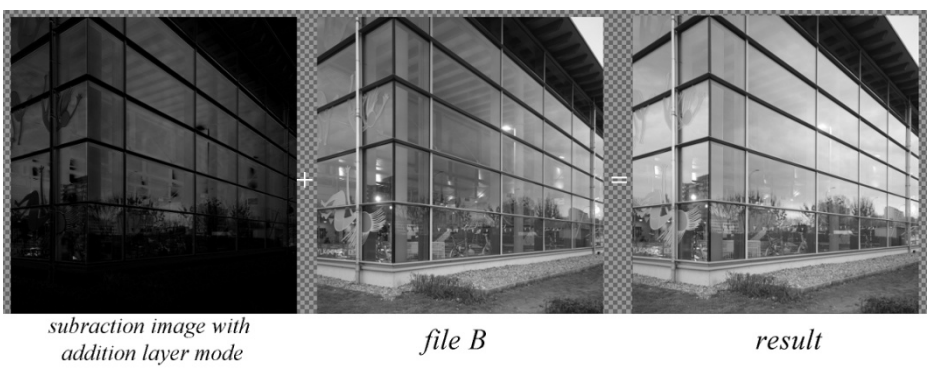


Fig. 4. Assigning transparency to *reflection overlay* by manipulating the Layer Mode of the *subtraction image* layer

This operation "reverses" the previously performed subtraction, but allows the researcher to independently manipulate the *reflection overlay*, which is on a separate layer (see Fig. 4). The layer could be switched on and off, its transparency could be adjusted, as well as the image transformations could be applied to this layer.

6.3 Using the Alpha Channel vs. Layer Mode Approach

As depicted in Fig. 5. the results of both approaches slightly differ with the latter method delivering a more intense image. This image more accurately reproduces the original *file A*, i.e. the one with virtual image. This could be the result of the difference in the algorithms of assigning transparency for each approach. The blending of images with alpha channel usually follows the algorithm: (RGB *file B*) + (RGB s*ubtraction image* * alpha); whereas in the layer mode approach, the algorithm is a simple RGB addition (RGB *file B*) + (RGB s*ubtraction image*).

Fig. 5. The difference in the final outcome between the alpha channel method and the layer mode method of assigning transparency to the *reflection overlay*

7 Modifications of the Transparent *Reflection Overlay*

The separation of the transparent *reflection overlay* allows the researcher to manipulate the virtual image independently of the other layer. The possible modifications of the transparent *reflection overlay* include: *duplication*, *multiplication with shift* (i.e. duplication with a change of location), barrel and pincushion *geometrical distortion,* and *tinting* and *blurring* of the virtual or real image.

- *Duplication* of the virtual image increases its luminance and makes it more visible than the real image. Adding another layer, with the transparent *reflection overlay,* causes a logarithmic increase in the luminance of the virtual image, which, with each subsequent layer, adds up until the maximum value of white (255, 255, 255) is reached.
- *Multiplication with shift* leads to an interesting perceptual phenomenon: depending on the value of the shift, part of the field of view seems out of focus. This simulates a real-life optical phenomenon, where two light-permeable panes are located in relatively close proximity and the two superimposed virtual images are not recognized by the observer as separate, but as one image that is out of focus. This cognitive phenomenon is quite extraordinary. Parts of the virtual image seem blurred, while the rest of the observer's field of view is in focus.
- *Geometrical distortions* allow for the simulation of the non-linear geometry of the reflecting surface. In the case of curved panes, the visible virtual image becomes the "reflection of the environment" – that is why the use of real-life photographs

was previously postulated. The extent of the distortions and the location of highlights (regions of higher luminance) can constitute an important cue for the human visual system. It seems that the human visual system "knows" the principles that shape the image on the light-permeable pane well enough and, "guided" by the location of highlights, "decodes" the 3-dimensional form of a smooth object. Those highlights also "affect observers' judgment of surface gloss" [5, p. 165], It seems reasonable to theorize that observation of the highlights results in the impression of "smoothness", which occurs at mid-level perception.

- *Tint* applied to the virtual image reproduces the perception of light-permeable materials that absorb certain wavelengths of light. The light can be absorbed during transmission – in which case the real image is tinted, or during reflection, in which case the virtual image is tinted. Both processes can be modeled using the above-mentioned method with consideration of the specific results of the con-fuse of luminous fluxes. A monochromatic tint overlaying the entire field of view makes the light-permeable surface clearly recognizable.
- *Blur* applied to the virtual image reproduces the perception of light-scattering materials which cause *(i)* blurring of the edges of objects located on the other side of the pane, and *(ii)* a decrease in contrast that is much higher than the one resulting from absorption.

8 Conclusions

In comparison with the previously proposed method of "pictorial image analysis", the new algorithm of *reflection overlay* utilizing the addition layer mode is much more precise in evaluating the influence of the virtual image on the perception of transparency in architecture and gives the researcher freedom to modify the conditions of the preformed tests. The new algorithm is more objective, less biased and less user-dependent. Provided that the conditions of the data acquisition are strictly followed the algorithm could be applied almost automatically.

Due to its relative simplicity (only basic equipment and software are required) and clear, easy-to-follow, step-by-step guidelines, the presented transparent *reflection overlay* method can be used by architects in assessing the results of their design.

References

1. Jaskowski, P.: Neuronauka poznawcza, jak mozg tworzy umysl. Wizja Press &IT, Warsaw (2009)
2. Bell, M.: Introduction... In: Bell, M., Kim, J. (eds.) Engineering Transparency: The Technical, Visual, and Spatial Effects of Glass. Princeton Architectural Press (2008)
3. Brzezicki, M.: The Role of Specular Reflection in the Perception of Transparent Surfaces – The Influence on User Safety. In: Harris, D. (ed.) EPCE 2013, Part I. LNCS, vol. 8019, pp. 189–196. Springer, Heidelberg (2013)
4. Abramoff, M., Magelhaes, P., Ram, S.: Image processing with ImageJ. Biophotonics International 11(7), 36–42 (2004)
5. Blake, A., Bulthoff, H.: Does the brain know the physics of specular reflection. Nature 343(6254), 165–168 (1990)

Analysis of Visual Performance during the Use of Mobile Devices While Walking

Jessica Conradi and Thomas Alexander

Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE
{Jessica.Conradi,Thomas.Alexander}@fkie.fraunhofer.de

Abstract. Mobile computers and smartphones are often used while their users are walking. From an ergonomic viewpoint, this requires a thorough design of the user interface. Although styleguides provide multiple recommendations there is little known about basic human factors' issues. This study provides recommendations for the visual design by analyzing the influence of walking on visual acuity with a mobile computer. N=22 volunteers participated in the experiment comparing visual acuity during standing, slow walking and fast walking. Additional conditions referred to indoor (treadmill) and outdoor (free walking) situations. The results show that walking speed has a highly significant influence on visual acuity. The results are independent of the indoor or outdoor condition. The decrease of visual acuity is similar to a row on a common eye chart. For compensating this decrease, letters and icons on a mobile device should be enlarged by about 20%.

Keywords: Dynamic visual acuity, DVA, smartphone, walking, mobile use, letter size.

1 Introduction

Technological improvements have let to small-sized displays of smartphones with increasing resolution and, thus, an increasing number of pixels per inch. Today's smartphones offer maximum resolutions beyond visual perception. Because of this it is possible to display miniaturized and undersized letters and icons, which cannot be perceived without extra efforts. Furthermore, small and light-weight mobile devices are frequently used concurrently with other activities. Users often text, retrieve information or check e-mails whilst walking. It is obvious that this might be dangerous, especially for pedestrians in situations with a lot of traffic. But it is always frustrating and increases workload if the visualization is hampered by too small letters or icons. Although their size might still be suitable for interaction while the users are sitting or standing, it is too small for a reliable interaction while walking and simultaneously paying attention to the environment.

During recent decades pixel size has minimized for hand-held mobile devices. While a PDA such as the Apple Newton H1000 was built with a display with 79.4 ppi in 1993, the Dell Axim X50v 2004 provided 216 ppi 10 years later. In 2010, an iPhone 4 or an iPod touch (4. Generation) included a display with 326 ppi. Recent

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 133–142, 2014.

devices as the HTC One possess up to 468 ppi. This development makes e.g. the display of full-HD Videos possible, but it also comes along with a miniaturization of elements of the user interface. Considering a common distance between user's eye and device, the resolution is sufficient to display letters and icons in a minuscule size beyond normal eyesight. This is important to remember when designing user interfaces for these devices.

Visual acuity is an individual trait which is measured following a standardized procedure. According to ISO 8596 [1], visual acuity is determined using Landolt C optotypes. A "Landolt broken ring" consists of a circular ring with a gap, therefore resembling a "C". The position of the gap is varied resulting in eight different varieties of the optotype. The participant states the location of the gap. In the following the size of the optotype is reduced until errors in the participant's responses exceed a predefined rate of errors. The size of the gap and the distance between the participant and optotype determines the angle taken as the measure for the visual acuity (Minimum Angle of Resolution, MAR). The logarithmized angle (logMAR) defines normal vision at 0.0 logMAR. Steps of 0.1 logMAR are identical to the rows on an eye chart.

This standard is based on static conditions and no individual movement is considered. Nevertheless, reading tasks often involve movement of either the reader or the object. This is usually the case during walking or driving. To consider the resulting effects, visual acuity is sub-divided into static visual acuity (SVA) and dynamic visual acuity (DVA). DVA is defined as the ability to discriminate the details of an object while there is relative movement between participant and the object [2]. DVA can be measured during voluntary ocular tracking of moving objects. Relative movement is induced by moving either the display or the participant's body or head [3, 4]. A test with moving objects is obtained by rotating optotypes of different sizes in the field of view. Rotation velocity is reduced until the participant perceives the optotype correctly. This way a threshold can be determined for each rotation velocity and optotype size, respectively [5]. With a static optotype, the participant rotates the head voluntarily at a specific angular velocity. The stimulus is displayed when the specific velocity is reached. The control of the experimental conditions is difficult for such a setup [6]. In another study the participants were rotated by a mechanism. Compared to the previous setup with a self-paced rotation, the rotated participants achieved less DVA [7]. Nevertheless, there is no standardized procedure to measure and to describe DVA by now.

An angular velocity of less than 2°/sec has no effect on the visual acuity [8]. The DAV decreases with higher angular velocities. The eye's tracking ability is exceeded at a speed of more than 50 °/sec, and discrimination of stimuli is impossible [3]. Further factors affecting DVA are for instance contrast [9] or personal traits like age, gender or experience in certain sports [6].

SVA and DVA correlate, whereas correlation fades with increasing angular velocity [2, 3]. Nevertheless, correlation is low. In order to assess individual performance in certain tasks, both, DVA as well as SVA, should be considered [10].

Measuring DVA includes just a small range of well-controlled motion. Dynamic visual tasks in every-day life hardly take place under such controlled conditions. Especially during walking, speed and direction of most parts of the body change

continuously. Although the head is usually stabilized during walking, it although moves at an amplitude of 5-9 cm, a speed of 0,25-0,35 m/s and a rotation of 5° ± 2,5° (Mean ± SD), which results in a maximal angular velocity of 30°/s ± 8°/s [11].

There are few studies about DVA while participants are walking. One of these compares SVA while standing with DVA while running on a treadmill at a speed of 6.4 km/h. Optotypes (numbers) were presented at a distance of 2 m. Results show that SVA is significantly lower than DVA while running [12]. In another study, the distance between user and object was analyzed as an additional factor. The participants performed a test with Landolt-rings displayed at a distance of 4 m and 50 cm respectively. For a distance of 4 m there was no significant difference between SVA and DVA. But for a distance of 50 cm the visual acuity decreased by 2.3 rows according to ISO 8596 [13]. Another study referred to the influence of walking with different speeds on legibility of normal text and pseudo text on a mobile phone. The results show that visual performance decreased with increasing speed (1.5 km/h, 3 km/h and self-paced speed of 3.4 - 4.5 km/h). Error rate increased and reading time decreased [14]. However, letter size was not considered.

In addition, the biomechanics of walking also effect visual acuity [12]. But because of multiple changing translational and angular speed of the hand-arm-shoulder system, the amount of influence on visual acuity is uncertain and varies [11]. Therefore, it is hard to calculate the effect precisely.

The focus of this study is to predict the actual change in visual acuity using a smartphone while walking. This is especially important because of increasing display resolution and undersized letters.

The baseline hypothesis following the rationale is that walking has a negative effect on visual acuity compared to standing. In addition, it is hypothesized that the effect increases with walking speed.

Studies focusing on walking are frequently carried out applying treadmills in laboratory setups. This allows for a strong control of environmental factors (e.g. light, walking speed, distraction) and other variables. Data collection is usually easier because of additional equipment for measuring and data storage. But a comparative study of free walking vs. walking on a treadmill reveals differences for certain joint kinematics and other temporal variables [15, 16]. According to an usability study the added value of conducting the evaluation in field additionally to the treadmill was found to be very little [17]. In a further study, reading comprehension and word search tasks were administered while walking free vs. treadmill, both inside a laboratory. Therefore, light conditions were identical. The authors found no influence of walking condition on performance, but some differences in subjective measures [18].

As a consequence, this study also considers and investigates a potential effect of the treadmill and other laboratory characteristics on visual acuity.

The purpose of this study is to determine and quantify the influence of walking on visual acuity in a laboratory setup on a treadmill and in an outside setup facilitating free walking.

2 Method

2.1 Participants

N=22 (13 male, 9 female) participants volunteered to take part in the experiment. They were aged 31,0 ± 5,4 (Mean ± SD). Their individual static visual acuity was measured using a standard vision screening instrument (Rodenstock R22, examination disc 119) according to ISO 8596 prior to the experiment [1]. All participants had normal or corrected to normal vision (LogMAR ≤ 0.0).

2.2 Apparatus

Well-established and standardized tests are available to measure visual acuity [1]. But for a more sophisticated analysis the single value for visual acuity of such standardized tests is insufficient. Instead, the psychometric function is analyzed. This function resembles the cumulative distribution function of a normal distribution. Characteristics are described by PSE (Point of Subjective Equality) and JND (Just Noticeable Difference). At PSE the number of correct stimuli equals the number of incorrect stimuli. JND describes the slope of the function and resembles standard deviation, but refers to 25% and 75 % recognition rate.

An adequate data basis is required to determine an individual psychometric function. It has to provide sufficient data especially in the range close to the PSE. An adaptive double-staircase-procedure was applied to cover this requirement. In this procedure the size of a stimulus depends on the individual performance at the preceding stimulus. If the first stimulus was determined correctly, the size of the following one is reduced. In the other case the size of the following stimulus is enlarged [19].

Furthermore, the procedure of the visual acuity according to ISO 8596 describes a SVA test. But this present study requires a test for DVA. Therefore the standard SVA test was adapted for the dynamic scenario. It is also based on the optotype Landolt-ring. Each stimulus was presented separately and followed by a screen which allowed for the selection of the correct ring (see Fig. 1). Presentation time was 1 sec.

The visual test was presented on an Apple iPod Touch, 4th generation, which provides a so-called “retina” 3.5“ multi-touch display with 326 ppi. This equals a pixel-size of 0,078 mm.

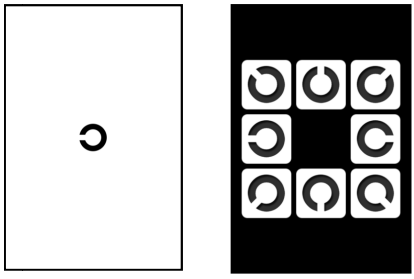


Fig. 1. Stimulus optotype Landolt-ring (left) and selecting interface (right)

The visual acuity refers to an angular value. This requires measuring both: stimulus size and distance between eye and stimulus. Therefore, position of the head and device were logged by an infrared motion tracking system (in the "inside" laboratory condition) and distance was calculated afterwards. For the outdoor condition, measuring system based on a visual marker on the device, a camera, and a subsequent pattern recognition was applied.

The participants kept a minimum distance between eye and smartphone of 45 cm. When the measured distance was below this limit, feedback was provided to the participant in order to correct the distance.

Walking speed is highly adaptable, but there are several attempts to identify "normal" or "preferred" walking speed. The average speed was found to be 4.92 - 5.04 km/h in adults, with an average of 4.62 km/h for females and 5.16 km/h for males [20]. In a meta-analysis regarding 41 studies with 23.111 participants the authors found an average of 4.9 – 5.2 km/h for males and 4.8 - 5 km/h for females. Therefore, we selected "normal" walking speed at 5 km/h and "slow" walking speed at 2.5 km/h.

The experiment was conducted in a laboratory for the condition "inside". The experimental apparatus included a treadmill, which facilitated walking speeds of 2.5 and 5 km/h (see Fig. 2, on the left). Illumination of experimental area matched ISO 8596 [1].

The condition "outdoor" was carried out in a straight, shady, quiet, tarred road in the vicinity of the institute (see Fig. 2, on the right). Weather conditions such as rain or bright sunshine were excluded. The participants practiced to keep the walking speed of constantly 2.5 and 5 km/h and achieved a precision of ± 0.2 km/h.

Fig. 2. Participant performing the experiment during the indoor/laboratory (left) and outdoor condition (right)

2.3 Design and Procedure

A 3 x 2 design with repeated measures on both factors was used for the study.

The first factor "walking" was varied in three levels: "standing" (0 km/h), "slow" walking (2.5 km/h) and "normal" walking (5 km/h). To exclude any sequence effects, order of conditions was permuted. Participants were assigned randomly to a permutation.

The second factor "environment" consisted of the conditions "indoor" and "outdoor". Because of weather, the condition "outdoor" was carried out two months later than "indoor".

Each experimental session started with a standard SVA test and a short introduction to the mobile DVA test. Subsequently, the participants were equipped with the distance measuring equipment. The following task consisted of fulfilling three double-staircase procedures (240 – 300 stimuli) per condition.

The psychometric function was determined for each participant in each condition. The characteristic parameters of the psychometric function, point of subjective equality (PSE) and just noticeable difference (JND), were calculated. They were considered as dependent variable in the following statistical analysis.

2.4 Statistical Analysis

The statistical distribution of all data sets was tested by a Kolomogorov-Smirnov test for normal distribution. All data showed normality. The three-level factor "walking" was tested by a Mauchly-test and no violations of sphericity occurred. Consequently, a 2x3 MANOVA with repeated measures on both factors was carried out. In case of significant differences, a pairwise comparison (Bonferroni corrections) followed for the three-level factor "walking". A significance level of 5% was used for the statistical analyses.

3 Results

The results of the statistical analysis show an influence of walking on visual acuity.

For PSE the smallest logMAR was found "standing"/"outdoor", while the highest was "fast"/"indoor". The values for both conditions are summarized in Table 1.

Table 1. Mean and standard deviation of PSE in logMAR

	indoor			**Outdoor**		
	standing	slow	normal	standing	slow	Normal
mean	-.2184	-.1466	-.1228	-.2287	-.1397	-.1402
SD	.08657	.08920	.11048	.08230	.07316	.07591

The repeated measures MANOVA revealed a highly significant influence of "walking" ($F_{(2.42)}$=37.12, p<0.01, η^2=0.639). "Environment" ($F_{(1.21)}$=0.341, p=0.566, η^2=0.016) as well as the interaction ($F_{(2.42)}$=0.655, p=0.525) had no effect on PSE.

A multiple comparison (Bonferroni correction) showed significant differences for the levels "standing" and "slow" (p<0.01) and "standing" and "fast" (p<0.01) but not for "slow" and "fast". The significant difference was 0.08 logMAR ("standing"/"slow") and 0.092 ("standing"/"fast") which almost equals a row on an eye chart. An illustration of the results is given in Fig. 3.

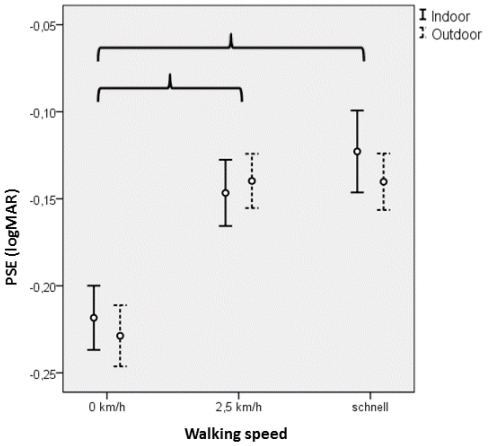


Fig. 3. Means and SD for PSE, parentheses indicate significant differences

The results for the measure of dispersion JND varied from 0.740 to 0.976, whereas the highest results occurred at "fast"/"outdoor". Means and SDs for all conditions are given in Table 2.

Table 2. Mean and standard deviation of JND in logMAR

	indoor			**outdoor**		
	standing	slow	normal	standing	slow	normal
mean	.0754	.0740	.0818	.0787	.0849	.0976
SD	.03357	.03120	.02689	.03601	.03746	.03443

The following repeated measures MANOVA showed a significant influence of the factor "walking" ($F_{(2.42)}$=3.837 p=0.029, η^2=0.155), but no influence for "environment" ($F_{(1.21)}$ =0.341, p=0.566, η^2=0.016) and no interaction respectively ($F_{(2.42)}$=0.535, p=0.589).

Bonferroni-corrected multiple comparison revealed a significant difference for "standing"/"fast" (p=0.034) as well as "slow"/"fast" (p=0.035). The differences amounted to 0.013 („standing"/"fast") and 0.010 („slow"/"fast"), respectively. Fig. 4 illustrates the statistical values.

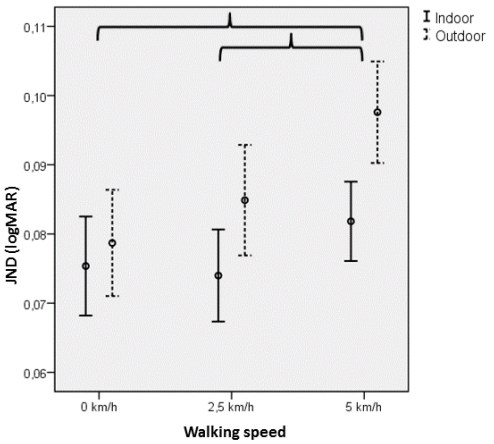


Fig. 4. Means and SD for JND, parentheses indicate significant differences

4 Discussion

This study proofs an influence of walking on dynamic visual acuity. Visual acuity is highest while users are standing and DVA is reduced while they are walking. This matches with the results of Hillmann und Bloomberg [12] and Peters und Bloomberg [13], respectively. In contrast to the results of Peters und Bloomberg who found a reduction of 2.3 rows of an eye chart, we found a reduction of one row only. This difference could be caused by the fact that Peters und Bloomberg used a stationary chart for their experiment, while participants held the smartphone in their hands in this study. Therefore, the participants were able to compensate head movements using their hand-arm-shoulder-system.

Another result is that only walking has an effect on visual acuity as opposed to standing. Walking speed does not affect PSE and, thus, DVA. The effect occurs even at low speeds. One conclusion is that speed reduction does not help to improve visual acuity. But there are other ways to compensate the loss in visual acuity. This is by either shortening the distance between eye and device or by adapting the size of letters and icons. It can be achieved by enlarging letters or icons by about one row of an eye chart, which equals to 20%.

JND was susceptible to walking speed. The slope of the psychometric function flattens in fast walking compared to slow walking. Faster walking also triggers incorrect detections of PSE-exceeding optotypes. This may result in an increasing number of errors. This effect can be reduced by adapting the walking speed and slow walking.

The environmental situation showed no effect on visual acuity. This corresponds to the findings of studies concerning performance measures [17, 18]. Other studies showed an influence of treadmill walking on physiological measures [15, 16]. However, this influence does not extend to visual acuity. It is concluded that our laboratory setting matches well with outside conditions and results are transferable. Nevertheless, in our outdoor setting environmental factors including light were

limited to a comparatively small range. Moreover, the outdoor setup was characterized by few or no additional distracting stimuli (i.e. pedestrians, cars, or other obstacles).

The results show that walking has a considerable influence on visual acuity for mobile devices. The change in visual acuity results into a reduction of one level on a typical test chart for visual acuity. The slope of psychometric function also flattens for faster walking. This also indicates a less stable performance in visual acuity and results into more errors. Consequently, the display of information has to be adapted or at least adaptable to the different use cases of the mobile device. The sizes of icons and letters should be increased by about 20% to compensate the loss in visual acuity caused by walking. This becomes more relevant if mobile devices are used in traffic situations with a lot of distracting environmental stimuli.

References

1. ISO 8596: Ophthalmic optics - Visual acuity testing - Standard optotype and its presentation. International Organization for Standardization, Geneva (2009)
2. Burg, A.: Visual acuity as measured by dynamic and static tests: a comparative evaluation. Journal of Applied Psychology 50(6), 460–466 (1966), doi:10.1037/h0023982
3. Miller, J.W., Ludvigh, E.J.: The effect of dynamic visual acuity. Survey of Ophthalmology (1), 83–116 (1962)
4. Ludvigh, E., Miller, J.W.: Study of Visual Acuity during the Ocular Pursuit of Moving Test Objects I Introduction. J. Opt. Soc. Am. 48(11), 799 (1958), doi:10.1364/JOSA.48.000799
5. Brown, B.: Resolution thresholds for moving targets at the fovea and in the peripheral retina. Vision Res. 12(2), 293–304 (1972)
6. Banks, P., Moore, L., Liu, C., Wu, B.: Dynamic visual acuity: a review. The South African Optometrist 63(2), 58–64 (2004)
7. Tian, J.-R., Shubayev, I., Demer, J.L.: Dynamic visual acuity during passive and self-generated transient head rotation in normal and unilaterally vestibulopathic humans. Exp. Brain Res. 142(4), 486–495 (2002), doi:10.1007/s00221-001-0959-7
8. Demer, J.L., Amjadi, F.: Dynamic visual acuity of normal subjects during vertical optotype and head motion. Invest. Ophthalmol. Vis. Sci. 34(6), 1894–1906 (1993)
9. Lit, A.: Visual Acuity. Annu. Rev. Psychol. 19(1), 27–54 (1968), doi:10.1146/annurev.ps.19.020168.000331
10. Lüder, A., Böckelmann, I.: Beurteilung des Zusammenhanges zwischen dem dynamischen Sehen und den Parametern statischer Visus sowie Kontrastempfindlichkeit. Praktische Arbeitsmedizin (21), 22–27 (2011)
11. Pozzo, T., Berthoz, A., Lefort, L.: Head stabilization during various locomotor tasks in humans. Exp. Brain Res. 82(1) (1990), doi:10.1007/BF00230842
12. Hillman, E.J., Bloomberg, J.J., McDonald, P.V., Cohen, H.S.: Dynamic visual acuity while walking in normals and labyrinthine-deficient patients. Journal of Vestibular Research 9(1), 49–57 (1999)
13. Peters, B.T., Bloomberg, J.J.: Dynamic visual acuity using “far” and “near” targets. Acta Otolaryngol. 125(4), 353–357 (2005), doi:10.1080/00016480410024631

14. Mustonen, T., Olkkonen, M., Hakkinen, J.: Examining mobile phone text legibility while walking. In: Dykstra-Erickson, E., Tscheligi, M. (eds.) Extended Abstracts of the 2004 Conference, Vienna, Austria, p. 1243 (2004), doi:10.1145/985921.986034
15. Alton, F., Baldey, L., Caplan, S., Morrissey, M.C.: A kinematic comparison of overground and treadmill walking. Clin. Biomech (Bristol, Avon) 13(6), 434–440 (1998)
16. Stolze, H., Kuhtz-Buschbeck, J.P., Mondwurf, C., Boczek-Funcke, A., Johnk, K., Deuschl, G., Illert, M.: Gait analysis during treadmill and overground locomotion in children and adults. Electroencephalogr. Clin. Neurophysiol. 105(6), 490–497 (1997)
17. Kjeldskov, J., Skov, M.B., Als, B.S., Høegh, R.T.: Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field, pp. 61–73 (2004)
18. Barnard, L., Yi, J.S., Jacko, J.A., Sears, A.: An empirical comparison of use-in-motion evaluation scenarios for mobile computing devices. International Journal of Human-Computer Studies 62(4), 487–520 (2005), doi:10.1016/j.ijhcs.2004.12.002
19. Levitt, H.: Transformed up-down methods in psychoacoustics. J. Acoust. Soc. Am. 49(2, suppl. 2), 467+ (1971)
20. Perry, J.: Gait analysis. Normal and pathological function. SLACK Inc., Thorofare (1992)

Model-Based Analysis of Two-Alternative Decision Errors in a Videopanorama-Based Remote Tower Work Position

Norbert Fürstenau, Monika Mittendorf, and Maik Friedrich

German Aerospace Center, Inst. of Flight Guidance, Braunschweig, Germany
Norbert.fuerstenau@dlr.de

Abstract. Initial analysis of a first Remote Control Tower (RTO) field test with an experimental videopanorama system [1] [2] under quasi operational conditions has shown performance deficits quantified by two-alternative aircraft maneuver discrimination tasks [3]. RTO-controller working position (CWP-) performance was compared with that one of the conventional tower-CWP with direct out-of-windows view by means of simultaneous aircraft maneuver observations at both operator positions, and it was quantified using discriminability d' and Bayes inference. Here we present an extended data analysis using nonparametric discriminability A and we discuss the RTO performance deficit in terms of the information processing (IP) theory of Hendy et al. [4]. As initial working hypothesis this leads to the concept of time pressure (TP) as one major source of the measured response errors. We expect the RTO-performance deficits to decrease with the introduction of certain automation features to reduce time pressure and improve the usability of the videopanorama system. A fit of the experimental data with a modified error vs. TP function provides some evidence in support of the IP/TP-hypothesis, however more specifically designed experiments are required for obtaining sufficient confidence.

Keywords: Remote Tower, videopanorama, field testing, flight maneuvers, two-alternative decisions, signal detection theory, information processing theory, time pressure.

1 Introduction

Since about ten years remote control of low traffic airports (Remote Tower Operation, RTO) has emerged as a new paradigm to reduce cost of air traffic control [1]. It was suggested that technology may remove the need for local control towers [5]. Controllers could visually supervise airports from remote locations by videolinks, allowing them to monitor many airports from a remote tower center (RTC) [2]. It is clear from controller interviews that usually numerous out-the-window visual features are used for control purposes [6]. In fact, these visual features go beyond those required by regulators and ANSP's (air navigation service providers) which typically include only aircraft detection, recognition, and identification [7]. Potentially important additional visual features identified by controllers in interviews involve subtle aircraft motion. In

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 143–154, 2014.

fact, the dynamic visual requirements for many aerospace tasks have been studied, but most attention has been paid to pilot vision (e.g. [8]). In this work we investigate a group of visual cues derived from flight maneuvers within the range of observability in the control zone. They might be indicative of aircraft status and pilots situational awareness which is important with the higher volume of VFR traffic in the vicinity of small airports.

These considerations led to the design of the present validation experiment within the DLR project RAiCe (Remote Airport traffic Control Center, 2008 – 2012). The field test was realized within a DLR - DFS (German ANSP) Remote Airport Cooperation. Specifically dual-choice decision tasks (the subset of "Safety related maneuvers" in [9]) were used for quantifying the performance difference between the standard control tower work environment (TWR-CWP) and the new RTO controller working position (RTO-CWP) based on objective measures from signal detection theory (SDT)[10] and Bayes inference [3]. Here we confirm these preliminary results by additional data evaluation using the nonparametric discriminability index A [11] and present a new model-based analysis in terms of the information processing/time pressure (IP/TP-) theory of Hendy et al.[4] for comparing the measured performance deficit of the RTO-CWP with the predictions of a theoretical error model.

Experimental methods are reviewed in section 2 followed by the results in section 3 (response times, Hit and False Alarm rates). Using these data in section 4 nonparametric discriminability coefficients are calculated and error rates are fitted with a IP/TP based model. We finish with a conclusion and outlook in section 5.

2 Methods

In what follows we review the experimental design with two-alternative decision tasks as part of the remote tower validation experiment and present additional details relevant for the IP-theory based analysis. Further details of the full passive shadow mode validation trial are reported in [9].

2.1 Participants

Eight tower controllers (ATCO's) from DFS were recruited as volunteer participants for the experiment. The average age was 30 (stdev 12) years with 10 (stdev. 10) years of work experience, and they came from different small and medium airports. They took part at the experiment during normal working hours and received no extra payment. They were divided into 4 experimental pairs for simultaneously staffing the control tower (TWR-CWP) and the RTO-CWP.

2.2 Experimental Environment and Conditions

The experiment was performed as passive shadow mode test under quasi operational conditions on the four days July 17 – 20 2012. The remote tower system used in the present experiment was located at the DFS-operated Erfurt-Weimar (EDDE) control tower. It was an improved version of the RTO-experimental testbed at Braunschweig

airport which was in use since 2004 for initial verification and validation trials [1][2]. Figures 1 show the sensor system and the RTO-CWP with 200° - videopanorama and operator console based on a reconstructed far view with five HD-format 40”-displays (892 x 504 mm, 1920x1080 pixel, pixel distance = 0.47 mm). A separate monitor (left table side) displays the pan-tilt zoom camera which is controlled via a pen-input interaction display with virtual joystick. Twelve preset viewing directions and four zoom factors Z = 2, 4, 8, 16 (viewing angles 26° - 3°) could be selected. Additional monitors include (middle row from right to left): flight plan data, approach radar and weather display. In contrast to the experimental TWR-CWP the RTO-CWP was on the TWR ground floor in a separate room without visual contact to the airfield. The TWR-CWP was located close to the operational ATCO, but they were instructed not to communicate with each other.

Fig. 1. Remote tower installation with 200°-panorama and pan-tilt zoom camera sensor system at the tower roof (left photo), and operators workplace with 40”-HD-format displays (right)

Pre-defined flight maneuvers were generated with a DLR DO228 twin turboprop engine test aircraft (D-CODE, length 15.03 m, body height x width 1.8 x 1.6 m, wing span 16.97 m, wheel diam. 0.65 m) performing aerodrome circling.

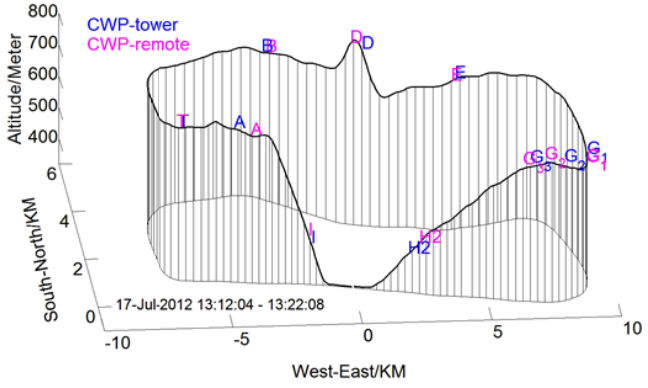


Fig. 2. DCODE trajectory measured with on-board satellite navigation. Letters indicate event positions with logged answers of TWR-CWP (blue) and RTO-CWP (red) operators to simultaneous task oriented on-line questioning. Distance between vertical lines = 5 s, projected to trajectory altitude minimum at ca. 350 m (sea level). Tower position (0, 0), height = 30 m.

The set of five well defined flight-maneuvers as stimuli for decision tasks at given positions within the EDDE control zone is indicated in Fig.2 with a 3D plot of the logged on board GPS trajectory. Trajectory minimum altitude represents a runway overflight at about 30 m above ground. The two types of maneuver-stimuli could be observed either visually-only (e.g. landing gear down) or visually and by radar (altitude change). During the experiment sometimes additional low volume normal traffic took place which now and then lead to delays in the traffic circle. Average duration of a full circle (= one run) was ca. 10 min yielding typically 140 min of experiment duration per participant pair for the nominally 14 full circles.

Radio communication between D-CODE pilots and flight engineer and the experimenter at the tower was realized with a separate research frequency in addition to the standard A/C–TWR radio channel. The available time for participant's responses to decision tasks was limited so that correct, incorrect and non-answers were possible.

2.3 Experimental Design and Task

Based on the fixed set of evaluated two-alternative events during a single circle (A, D, G_1, G_2, $H_{1,2,3}$) the concrete event situation (stimulus alternative S1 = maneuver, S2 = no maneuver) for decision making were switched statistically between runs according to two mirrored scenarios with different task sequences. In this way during the 14 circles per experiment / participant for each event about 7 answers per event alternative and per participant were obtained for the analysis (TWR and RTO CWP condition as independent variables). The within subjects design (TWR vs. RTO-CWP) yields about N = 56 answers and non-answers for averaging with the 8 participants.

During one aerodrome circling the two participants at the TWR and RTO-CWP had to simultaneously answer 19 different types of questions concerning the D-CODE maneuvers (events), object detection, and weather status. For analysis with discriminability index A and IP/TP theory based error model the following subset of 5 of the 9 event related questions is evaluated with regard to hit- and false alarm rates (in brackets: maximum response time Ta = interval until next task / question): 1. Does A/C perform repeated bank angle changes? (event position A; Ta = 20 s), 2. altitude variation ? (by 300 ft, event pos. D; 20 s), 3. landing light-off? Report status, event pos. G_1; 180 sec: switching-off not observable), 4. A/C on glide path?, event pos. G_2; 90 s; 5. Landing gear-down?; Report during final approach;event pos. $H_{1,2,3}$, distance 1.5, 1, 0.5 km; 10 s). A subjective certainty rating on a 5-point scale was not evaluated for the present analysis.

Every pair of participants had to complete two experimental trials. For the first trial (duration approximately 140 min) in the morning till noon, the participants were randomly assigned to one of the two CWP's. Positions were changed for the second trial in the afternoon. The number of airport circles corresponding to the number of decisions for each specific event was between 10 and 14. With two alternative stimuli (S_1, S_2) per task each pair of participants was exposed to 5 – 7 stimuli of the same kind in random succession for each event under TWR and the same number under RTO conditions. The present data analysis is focused on deriving objective measures for the two-alternative decision tasks. Additional data evaluation was presented in a parallel

paper [9] addressing performance (answers given, response times, and sources of information) and subjective measures (debriefing, questionnaires).

3 Results

The response matrices of conditional probabilities $p(y|S_1)$ = hit rate H, $p(n|S_1)$ = misses M, $p(n|S_2)$ = correct rejections CR, $p(y|S_2)$ = false alarms FA, for the two alternative situations (stimuli), S1, S2, structure the results of each of the five events. Because participant's responses to event related questions were allowed to be positive, negative, and non-answers (no decision during the available time Ta), we analyse two types of response matrices: a) (optimistic) neglecting non-answers, b) (pessimistic) interpreting non-answers as false decisions (M or FA). In this way we obtain for each of the decision tasks an optimistic and a pessimistic estimate with regard to decision errors. The percentage correct analysis in [9] and the preliminary SDT and Bayes inference analysis [3] had shown that neglection of non-answers suggested no significant performance difference between TWR-CWP and RTO-CWP. The interpretation of the non-answers as erroneous responses appears to be justified due to increased uncertainty about the correct answer resulting in hesitation to respond at all because tower controllers work ethics requires decision making with high certainty. Figure 3 shows the statistics of non-answers, separated for the TWR-CWP and RTO-CWP condition.

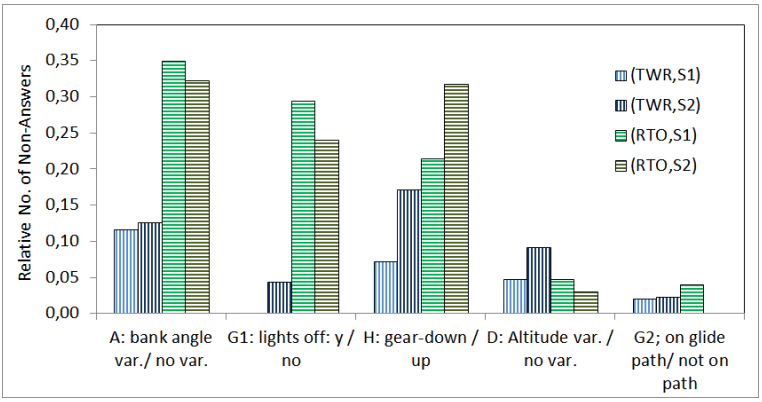


Fig. 3. Relative number of non-answers for the five analyzed decision tasks, separated for the two conditions TWR-CWP(left two columns, blue, vertical lines), RTO-CWP(right columns, green, horizontal lines), normalized with regard to the two respective alternative situations S1 (flight maneuver / stimulus, light colour), S2 (no flight maneuver / stimulus, dark colour))

Within the theoretical framework of SDT the two alternative stimuli S_1, S_2 for each event define independent statistical variables. Each set of decisions of a single subject for the 14 aerodrome circles with one of the events A, D, G_1, G_2, H represents a sample of the randomly presented S_1- and S_2-alternatives. For calculation of (parametric) discriminability d' the subjective responses are assumed to be drawn from

independent equal variance Gaussian ($\mu_{1,2}$, σ) densities for familiarity with situations S_1 and S_2 [10]. Any discriminability difference between TWR and RTO may be quantified by corresponding coefficients d' = $\mu_1 - \mu_2$ = z(H) – z(FA), and subjective decision bias (criterion) c = 0.5(z(H) + z(FA)), with z() = z-score as calculated from the inverse cumulative densities. This SDT-analysis together with Bayes inference on risk of false decision was provided in [3] for the events A, D, G1, H. In section 4.2 we will confirm these preliminary results with an additional analysis using the non-parametric discriminability index A [11] (independent of Gaussian assumption).

Table 1 lists the measured hit and false alarm rates (± standard deviations derived from binomial distributions) for the five events to be analysed with respect to A. In addition to H and FA, M = 1-H is required for calculating the total number of errors to be compared with a formal error model in section 4.3

Table 1. Measured hit and false alarm rates (H = p(y|S1), FA = p(y|S2), ± stddev from Binomial distribution according to [10]) for five events and two conditions (TWR, RTO-CWP) with a) non-answers excluded and b) non-answers added to error rates FA and M. Ta = available decision time, Tr required average decision time with stderror of mean / seconds.

Event with Alternatives S_1 / S_2 (Ta/s)	Tr / s ± stderr	CWP	a) Non-answers excluded		b) Non-answers included	
			p(y|S_1)	p(y|S_2)	p(y|S_1)	p(y|S_2)
A: bank angle var.: y / n (20)	13.8±1.7	TWR	0.92±.04	0.08±.04	0.81±.06	0.20±.05
	14.0±1.1	RTO	0.93±.05	0.11±.05	0.60±.07	0.39±.07
D: Altitude var.: y / n (20)	8.8 ±1.4	TWR	0.80±.06	0.03±.03	0.77±.06	0.12±.06
	12.4±1.5	RTO	0.73±.07	0.03±.03	0.70±.07	0.06±.04
G1: lights off: y / n (180)	27.0±6.6	TWR	0.94±.04	0.25±.07	0.94±.04	0.28±.07
	95.4±7.4	RTO	0.92±.06	0.63±.08	0.65±.08	0.72±.07
G2: Glidepath y/n (90)	21.6±6.4	TWR	0.90±.04	0.32±.07	0.88±.05	0.33±.07
	34.2±8.1	RTO	0.92±.04	0.22±.06	0.88±.05	0.22±.06
H: gear-down: y / n (10)	8.1±0.9	TWR	0.98±.02	0.06±.04	0.91±.04	0.22±.06
	9.2±0.5	RTO	0.98±.02	0.07±.05	0.77±.06	0.37±.08

Comparing the measured hit and false alarm rates for all five events under TWR and RTO conditions with non-answers not considered (optimistic case a): left two data columns), the RTO-CWP exhibits no significant difference as compared to the TWR-CWP. If however, the non-answers are interpreted as erroneous responses and correspondingly attributed to rates FA and M (pessimistic case b): right two data columns), significant differences TWR vs. RTO are obtained (smaller H(RTO), larger FA(RTO)) for event/task A (bank angle variation?), H (gear down?), G1 (lights off?), whereas for event/tasks D and G2 responses again exhibit no significant difference. The latter two tasks reflect the fact that altitude information could be read directly from the radar display and operators were free to select their appropriate information source. An extremely high FA difference TWR vs. RTO is observed for both case a) and b) for the "lights-off" event which is reflected also in a large difference of decision distance (correlated with response time).

4 Data Analysis and Discussion

4.1 Technical Limitations

Technical parameters of the reconstructed far view with videopanorama and PTZ [1][2] leads to predictions concerning performance differences under the two conditions TWR and RTO-CWP. The measured performance also depends on the usage of the different available information sources, in particular videopanorama, PTZ, and approach radar. The visibility limitations of the videopanorama are quantified by the modulation transfer characteristic (MTF), with the digital (pixel) camera resolution providing the basic limit (Nyquist criterion) for detectable objects and maneuvers: angular resolution was estimated as $\delta\alpha \approx 2$ arc min $\approx$ 1/30° $\approx$ 0.6 m object size / km distance per pixel under maximum visibility and contrast (about half as good as the human eye (1 arcmin)). Reduced contrast of course reduces the discriminability according to the MTF and the question arises how the discriminability difference TWR vs. RTO-CWP is affected. The gear-down situation at positions H1- H3 with wheel diameter 0.65 m, e.g. can certainly not be detected before the wheel occupies, say, 4 pixels which for the 40" display (0.55 mm pixel size) means a viewing angle of ca 1 mm/2 m $\approx$ 0. 5 mrad corresponding to the visual resolution of the eye (1 arcmin) under optimum contrast. This estimate results in a panorama based gear-down detectability distance of < 500 m. It means that under RTO conditions this task requires usage of PTZ in any case for enabeling a decision. The same argument is valid for the detection of bank angle changes at position A following the overflight of the runway because it requires optical resolution of the A/C-wings. The "lights-off?"-decision (G1) has a somewhat different character because in situation S_1 (lights-off, answer "yes" = hit) observers usually wait until they actually detect the A/C whereas situation S_2 can be recognized at a larger A/C distance due to the higher contrast ratio of landing-light-on/background luminance.

4.2 Discriminability of Aircraft Maneuvers during Aerodrome Circling

Based on the extended set of data as compared to [3], the focus here is on quantification of the discriminability deficit of the RTO-CWP by means of the nonparametric sensitivity index A with corrected algorithms [11], and the derivation of initial evidence for the IP/TP hypothesis [4] as formal framework for explaining the measured performance decrease. In [3] the (H, FA)-data of table 1 (without G2) were analysed using parametric discriminability d' and Bayes inference (see section 3). With decisions based on visual observation using videopanorama and PTZ, both SDT and Bayes analysis showed consistently a significantly reduced discriminability for the three maneuvers A, G1, H, but not for altitude change D where radar provided the required information. Due to the d' dependency on Gaussian distribution parameters we test here the reliability of the preliminary results with the nonparametric discriminability parameter A [11] which is calculated directly from H, FA. A is the average area under the minimum and maximum area proper ROC-isosensitivity curves (constant d', [3][10]) and varies between 0.5 (d' = 0) and 1 (lim d'-> ∞) . Figure 4 (right)

depicts for analysis of case b) the A-values of the five tasks at A, D, G1, G2, H, for the two conditions TWR-CWP, RTO-CWP. Fig.4(left) shows one example of (A, b)-parametrized isopleths determined by the two TWR and RTO-CWP datapoints.

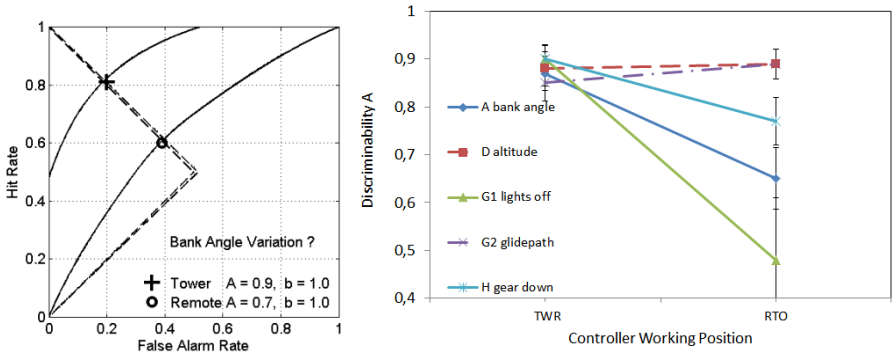


Fig. 4. Left: Isosensitivity curves (TWR, RTO) for maneuver A (solid lines, case b)-A-isopleths) and decision bias (dashed, b-isopleths). Right: A as calculated according to [11] from H and FA rates in table 1 for case b): non-answers := false answers. D and G2 (dash-dotted lines connecting TWR – RTO data): decisions about altitude (variations). A, G1, H = visual-only information (solid lines). Error bars = std. dev. based on binomial distribution [10].

The example (A, b)-isopleths for maneuver A (Fig. 4, left) shows zero decision bias (b = 1), however a significant discriminability decrease for RTO-CWP (circle = data average; minimum A-isopleth = 0.5 = positive diagonal). In agreement with table 1 and the d' results in [3], discriminability indices A in Fig.4(right) exhibit no significant difference between TWR and RTO-CWP conditions for events D, G2 (event sub-set with altitude stimulus; altitude information provided by radar), whereas the A-decrease for the visual-only subset {A, G1, H} is evident. Moreover even a reduction of the number of erroneous decisions by attributing a 50% chance to non-answers to be correct instead of assuming 100% wrong answers) leaves the RTO-performance decrease for visual-only tasks significant. The drop to chance level of RTO-CWP discriminability for case G1 is attributed to the RTO-resolution and contrast deficit which prohibits recognition of A/C even with lights on for short response times Tr: when participants at RTO-CWP after task initialization had waited some 10 s or so without recognizing landing lights they often simply guessed lights to be off or gave no answer, contributing to FA-errors.

4.3 The Information Processing / Time Pressure Hypothesis

In order to determine appropriate solutions for rising the RTO-CWP performance to at least the level of the TWR-CWP we have to find explanations for the measured discriminability deficits. The RTO-CWP performance for decision making using visual information only should be at least as good as that one based on radar used for the altitude related decisions (Fig. 4) so that users can be certain that replacement of the out-of-windows view has a potential of even improving their work condition.

A (algorithmically) simple theoretical model with some potential for explaining observed performance differences quantified in terms of decision-error probability, is based on the perceptual control/information processing theory (PCT/IP) of Hendy et al. [4]. Because our experiment was not initially designed for an application of this theory we can only expect a first impression on the relevance of the corresponding assumptions. The core idea is to formalize the information processed (as part of the total information required for a correct answer: Br / bits) as function of time pressure TP. TP is the ratio of required time Tr (to acquire Br) and the available time Ta: TP = Tr/Ta. Assuming constant cognitive processing rate (channel capacity C: Tr = Br/C) the rate of information processing demanded RID is related to TP via TP = RID/C, with RID = Br/Ta. Hendy et al. [4] derived simple algorithms for modeling dependent variables like operator workload (OWL), success ratio, and number of errors as function of TP. For the latter they suggested an exponential dependency for the increase of decision errors with TP, where TP increases linearly with the number N of objects to be analysed (in our case N = 1): TP = t_0(1 + b1 N)/Ta, and t_0 = minimal decision time for N = 0. For error probabilities we modify Hendy's algorithm in order to use our maximum error probability p_{err} = 0.5 = p_{max} (guessing, no information available) as boundary condition. Keeping the original assumption that errors start to grow exponentially with TP but then level off at p_{max} we arrive at a logistic function with threshold and sensitivity parameters as one possible model:

$$p_{err} = 0.5\left(1 + exp\left\{-\left(\frac{TP-\mu}{\beta}\right)\right\}\right)^{-1} \quad (1)$$

μ ($0 \leq \mu \leq 1$) models the threshold where the observer starts shedding most information due to increasing workload (stress due to TP increase). It fulfills the conditions that lim(TP >> μ) p_{err} → 0.5 and lim(TP → 0) p_{err} → 0. The latter condition is fulfilled as long as μ/β >> 1, i.e. steep slope (= error sensitivity $dp_{err}/dTP = 1/2\beta$ at TP = μ and/or large threshold). Figure 5 shows the results of nonlinear fitting of the respective two data points p_{err}(TP) at TP(Tr(TWR)), TP(Tr(RTO)) with the two boundary conditions (p_{err}(TP → 0, TP → ∞) using model-equation (1) for the three visual-only tasks. For characterising the experimental results in terms of (μ, β) we have to use the total number of errors for the full set (n(S1)+n(S2)) of trials per subject instead of the conditional probabilities, misses and false alarm rates M=1-H, FA: p_{err} = (n1 M + n2 FA)/(n1 + n2) as used for the discriminability calculation.

The results indicate the principal applicability of the logistic error model because all three cases yield reasonable threshold (μ < TP = 1) and error sensitivity parameters $1/\beta$. The RTO-performance deficit always seems to correlate with some kind of time pressure. According to IP-theory decision errors should increase significantly due to increasing stress when Tr approaches Ta and to shedding of information when Tr > Ta (Tr/Ta > 1). This is reflected by our results only for event H (gear down) with the shortest Ta = 10 s. Variation of threshold μ with event(stimulus) can be explained by the fact that the three specific events provide quite different stimulus conditions for the decision making as described in section 3.The fact that only for the gear-down task an approximately exponential increase of errors at TP ≈ 1 is observed according to [4] with $\mu \approx 1$ whereas a sensitive threshold behavior at lower μ is suggested for

tasks A, G1, indicates at least one more performance limiting factor besides time pressure, such as PTZ-camera contrast/resolution and operator training. For lights-off decision the RTO-HMI contrast deficit should play a major role: the average response appears completely at random. Nevertheless also in this case a long waiting time after beginning to gather visual evidence might lead to increasing stress due to uncertainty.

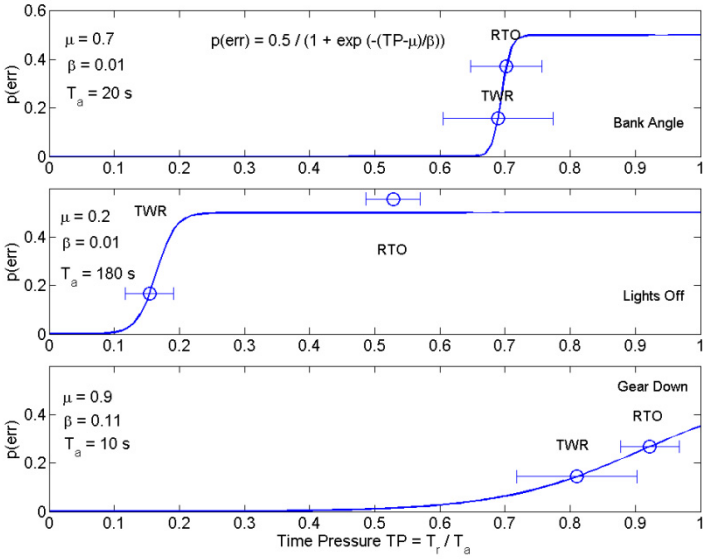


Fig. 5. Decision error probabilities for TWR and RTO-CWP vs. time pressure TP (±stderr of mean, n = n(error)+n(correct) = 80 …100) for tasks where visual / PTZ-information was used for decision making. Standard errors of p(error) are smaller than the circles of data points. Logistic error model (equ. (1)) derived from IP/TP-theory [4] for fitting p_{err}(TP).

5 Conclusion

The present analysis of two-alternative decision making with safety related aircraft maneuvers confirms the previously reported [3] explanation of an observed discrepancy in the percentage correct analysis (p_c, neglecting non-answers) [9] of the corresponding observation data, as compared to the subjective success criteria. The perceived safety was rated as insufficient by participants which agrees with the objective data of the present analysis and [3]. Neglecting non-decisions during simultaneous decision making at TWR- and RTO-CWP yields mostly no significant difference of discriminability (i.e. suggesting sufficient RTO performance) whereas the interpretation of non-decisions as false responses (misses or false alarms) leads to significant error increase under RTO as compared to TWR conditions and correspondingly reduced A and d'. The results indicate a usability deficit of the RTO-HMI (videopanorama and PTZ) due to time pressure as one possible reason. Data analysis with a modified version of the Hendy et al. information processing / time pressure

theory (IP/TP) [4] indicates additional origins of performance decrease due to threshold behavior of decision errors significantly below the TP = 1 value. It is expected that increased automation (e.g. automatic PTZ-object tracking and data fusion with approach radar) will increase usability, and in combination with improved operator training could solve the performance problem. However further experiments are required for clarifying the role of time pressure and validating the effect of a higher level of automation system. They are preferably realized as human-in-the loop simulations with appropriate design for time pressure variation, and forced choice tasks for avoiding non-answers. Because of the significant effort required for the HITL-experiments and field tests, the initial results of the IP/TP-model suggest as intermediate step computer simulations for preparing corresponding HITL- and field experiments. For this purpose the commercial tool IPME (Integrated Performance Modeling Environment [12]) appears useful which integrates the PCT/IP-based approach together with a resource based theory so that by means of simulations it would allow for further clarification of the influence of different performance shaping functions.

Acknowledgement. We are indepted to DFS personnel N. Becker, T. Heeb, P. Distelkamp, and S. Axt for cooperation during preparation of the experiment. Many thanks are due to C. Möhlenbrink and A. Papenfuß for support and performing online interviews during the exercises. Markus Schmidt and Michael Rudolph were responsible for the technical setup and RTO-software of the experiment. A. Grüttemann was responsible for the onboard data acquisition and served as flight engineer. We acknowledge the support of the DLR flight experiments department and in particular cooperation with pilots G. Mitscher and P. Bergmann.

References

[1] Schmidt, M., Rudolph, M., Werther, B., Möhlenbrink, C., Fürstenau, N.: Development of an Augmented Vision Videopanorama Human-Machine Interface for Remote Airport Tower Operation. In: Smith, M.J., Salvendy, G. (eds.) Human Interface, Part II, HCII 2007. LNCS, vol. 4558, pp. 1119–1128. Springer, Heidelberg (2007)

[2] Fürstenau, N., Schmidt, M., Rudolph, M., Möhlenbrink, C., Papenfuß, A., Kaltenhäuser, S.: Steps towards the Virtual Tower: Remote Airport Traffic Control Center (RAiCe). In: Proc. EIWAC 2009, ENRI Int. Workshop on ATM & CNS, Tokyo, March 5-6, pp. 67–76 (2009)

[3] Fürstenau, N., Friedrich, M., Mittendorf, M., Schmidt, M., Rudolph, M.: Discriminability of Flight Maneuvers and Risk of False Decisions Derived from Dual Choice Decision Errors in a Videopanorama-Based Remote Tower Work Position. In: Harris, D. (ed.) EPCE/HCII 2013, Part II. LNCS (LNAI), vol. 8020, pp. 105–114. Springer, Heidelberg (2013)

[4] Hendy, K.C., Jianquiao, L., Milgram, P.: Combining Time and Intensity Effects in Assessing Operator Information-Processing Load. Human Factors 39(1), 30–47 (1997)

[5] Hannon, D., Lee, J., Geyer, T.M., Sheridan, T., Francis, M., Woods, S., Malonson, M.: Feasibility evaluation of a staffed virtual tower. The Journal of Air Traffic Control, 27–39 (Winter 2008)

[6] Ellis, S.R., Liston, D.B.: Visual features involving motion seen from airport control towers. In: Proc. 11th IFAC/-IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, Valenciennes, France, September 31-October 3 (2010)
[7] Van Schaik, F.J., Lindqvist, G., Roessingh, H.J.M.: Assessment of visual cues by tower controllers. In: Proc. 11th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Valenciennes, France, August 31-September 3 (2010)
[8] Watson, A.B., Ramirez, C.V., Salud, E.: Predicting visibility of aircraft. PLoS One 4(5), e5594 (Published online may 20, 2009), doi:10.1371/journal.pone.0005594
[9] Friedrich, M., Möhlenbrink, C.: Which data provide the best insight? A field trial for validating a remote tower operation concept. In: Proc. 10th USA/Europe Air Traffic Management Research and Development Seminar (ATM 2013), June 10-13 (2013)
[10] MacMillan, N.A., Creelman, C.D.: Detection Theory. Psychology Press, Taylor and Francis, New York (2005)
[11] Zhang, J., Mueller, S.T.: A note on ROC analysis and non-parametric estimate of sensitivity. Psychometrica 70(1), 203–212 (2005)
[12] Fowles-Winkler, A.M.: Modelling with the Integrated Performance Modelling Environment (IPME). In: Proc. 15th European Simulation Symposium, SCS European Council (2003) ISBN 3-936150-29-X

Dynamic Perceptual Objects

Dennis J. Folds and Stuart Michelson

Georgia Tech Research Institute, Georgia Institute of Technology
Atlanta, Georgia USA
dennis.folds@gtri.gatech.edu

Abstract. A perceptual object is created when an observer perceives a single "thing" even though it is comprised of separately perceptible components. A perceptual object has permanence across changes in position and, within limits, changes in the arrangement or composition of the constituent parts. The present research is an examination of the emergence of perceptual objects solely from the dynamics of data presentation. Ten participants viewed presentations of dot patterns that varied in persistence, color, and opacity. Half the presentations were set to have parameters optimized to promote object perception. The same data were also presented with other (non-optimized) settings. Participants correctly detected about 95% of the targets presented with optimized settings and less than 5% of the same targets with non-optimized settings. There were very few false alarms. Participants perceived a unitary object that was hopping from place to place on display despite changes in speed, direction, or color.

Keywords: perceptual objects, beta motion, user interface, big data.

1 Introduction

The human visual perception system readily organizes the visual field into objects, and actively constructs missing attributes needed to complete those objects. Early Gestalt theorists noted that people tend to perceive the whole rather than the parts, and readily distinguish a figure from the ground [1]. The perceived object retains its identity across changes in perspective created by motion of the object or the observer, or both. Although the change in perspective may create significant distortion in the image that falls on the retina, the unity of the object is preserved. This perceptual capability emerges in infancy [2], and degradation of the capability may indicate a neurological disorder [3].

Object attributes that are perceived include direction and magnitude of motion. Perception of object identity is retained across changes in motion, spin, or rotation. These tendencies can be readily demonstrated with common everyday objects in the real world (e.g., cups and saucers), and with simple geometrical shapes presented on a display.

Given this robust tendency to organize individual components into higher-order objects, user interfaces may be designed to promote perception of objects even though only the individual components are plotted on a display. There has been considerable

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 155–163, 2014.

interest in discovering the principles that govern object perception, to provide guidance to display designers regarding effective graphical displays [4]. Creating a reliable perception of an object requires that the arrangement of the components matches human perceptual tendencies. It is particularly useful if object perception is intuitive for the observer. For example, a circle, two dots, and an arc, arranged within certain constraints, are perceived as the common "smiley face", not as four separate components. Outside those constraints, no face is perceived.

Matching human perceptual tendencies is important for visualization techniques to be effective. Visualization techniques should attempt to amplify cognition – not strain it [5]. One desirable property of a visualization technique (and perhaps a measure of its utility) is whether it promotes a rapid, intuitive understanding of relationships in the data in comparison to a tabular form [6]. The value of intuitive visualization is increased when dealing with large data sets, where inspecting a tabular representation of the data might not be practical. Visualization techniques that promote object perception could greatly increase understanding of relationships and trends in the data while greatly reducing the time required to achieve that understanding.

Dynamic elements in a perceptual object may convey information. For example, a gradual change in the orientation of the arc that comprises the mouth of the smiley face can convey emotion change from happy to neutral to unhappy ("frowny face"). In the present research, we examine perceptual objects that arise solely from the dynamic properties of components.

Presentation of images in rapid succession is the basis for motion pictures. The resulting perception of apparent motion, even though each image is a still frame, is called the phi phenomenon [7]. Motion is reliably perceived as long as the frequency (in frames per second) is sufficiently high, and the continuity of objects across frame is sufficiently consistent. The resulting perceptual experience is quite like the real world. If playback speed is too slow, jerkiness or disjointed movement may be perceived. If slower still, the presentation will be perceived as successive still frames.

A related phenomenon is beta motion [8], which is often created by on/off patterns of points of light (such as on a neon advertisement sign) in which adjacent elements are turned on or off in rapid succession, creating the perception of a moving ripple or wave proceeding across the array of elements. With beta motion, the perceptual object is an anomaly in the visual field that is moving systematically. It does not necessarily have a perceived identity.

In the present research, we demonstrate the emergence of dynamic perceptual objects in the presence of random dots that appear and eventually fade (or disappear completely) from a display. This perception is related to beta motion, but does not match the usual definition of beta motion. During the design of a fast replay capability for a sensor system, we noted the occasional emergence of beta motion (or at least something akin to it) when the replay was set at just the right speed and persistence to just the right level. At substantially slower or faster replay speeds the perception of motion did not occur. With no persistence or with permanent persistence, the perception was unlikely to occur. A given set of playback parameters (speed and persistence) did not create the effect for all data sets, because those sets were collected on different time scales. Clearly, the important parameters were related to the rate at which adjacent areas on the display (and consequently, on the retina) changed states.

As an analogy, consider the playback of weather radar data. If the playback speed and other parameters are within certain bounds, an observer can readily perceive the movement of a storm cloud or a weather front. For this perception to occur reliably, the playback speed (expressed as a multiple of real time) would be different for slow-moving clouds versus faster moving clouds. Data depicting slower moving clouds would need a faster playback speed, compared to faster moving clouds, to create an accurate perception of the speed and direction of movement of the cloud. Moreover, the playback speed would also vary for different range scales of the map. More zoomed-in range scales could be viewed at a slower playback speed than more zoomed-out range scales. What matters is the rate at which adjacent areas of the display change state. There is a range of values over which accurate perception is likely, but outside those ranges, accurate perception becomes less likely, and potentially unlikely. Optimizing the display presentation to promote accurate perception of object motion requires consideration of the rate of change of position of the object. The purpose of the present research was to demonstrate that optimized settings for presentation of dynamic data promote intuitive perception of moving objects, compared to non-optimized settings.

2 Method

2.1 Participants

Participants were five male and five female volunteers, ranging in age from 18 to 49. All had normal or corrected-to-normal visual acuity. All participants were friends and acquaintances of the investigators.

2.2 Equipment and Simulated Data

Ten synthetic data sets were created from a software simulation. Three were used in engineering development but not in the experiment. Two were used for training and the other five were used in evaluation. In each set, from one to three targets were present, with a random starting position in an arbitrarily defined x-y plane. At each time step in the simulation, one quadrant of the plane was sampled, and approximately 75 to 125 random false contact reports were generated, randomly located in the quadrant. On successive time steps, a different quadrant was sampled. When a target was present in the quadrant, a true contact report was generated at the target location. Thus for every target contact there were about 300 - 500 false contact reports. Each target was scripted to follow a designated path. Across data sets, no path of movement was replicated. Some targets left the plane before the simulation concluded. Some stopped altogether but remained in the plane; others paused and resumed motion. Some targets moved in a straight line, but most changed direction at least once; some changed often. Figure 1 depicts the pattern of motion for the two training and five evaluation trials.

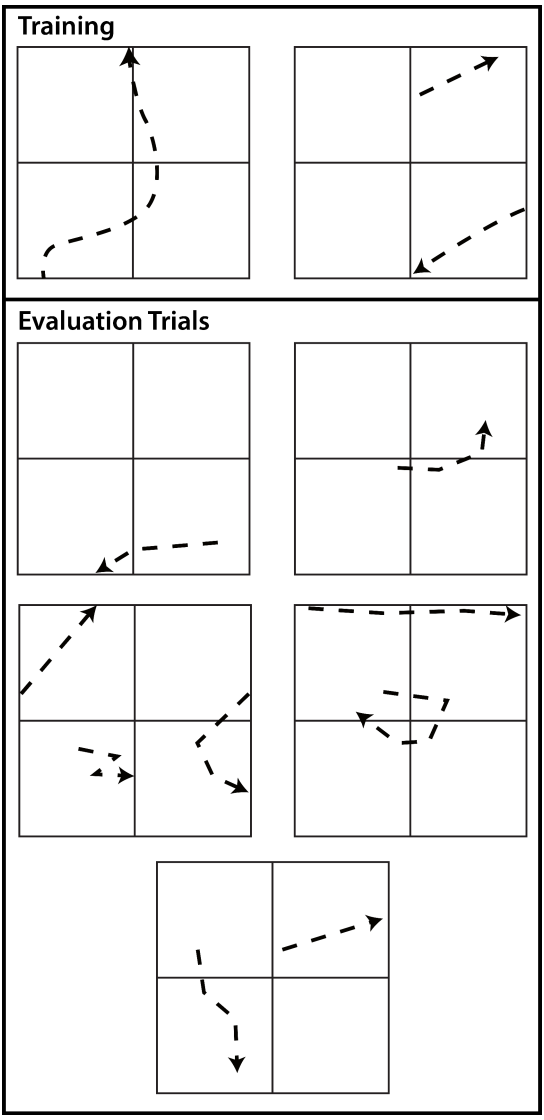


Fig. 1. Target motion in two training and five test trials

The playback software was configured with a designated set of playback parameters as shown in Table 1. Each of the five data sets was played back with two different parameter sets, one intended to promote object perception (optimized, or "good" parameters), or not to do so (non-optimized, or "bad" parameters). The notation in Table 1 shows the data set (4, 5, 8, 9, or 10) and the good (G) and bad (B) parameters. The playback software was allowed to loop (repeat) the playback three times, and the computer screen video was captured and stored, to provide a uniform playback for presentation to participants. All contacts were plotted as small colored dots on the

display. The target color was always either magenta (for the entire trial) or red and blue (switching color one or more times during the trial). Distractors (false contacts) were always the same color(s) as the target, and on some trials, there were also green distractors. The target was never green. Persistence was set to control how long a given contact dot remained on the screen independent of the playback speed. For the experiment, playback speed was held constant at approximately 6 Hz, which resulted in an update of target position at a rate of about 1.5 Hz, given the quadrant stepping procedure employed. At this playback rate, each playback loop was approximately 15 s in duration. The captured video files, with three loops, were about 45 s in duration.

Table 1. Playback parameters for evaluation trials

Trial ID	Target Color	Distractor Color	Persistence
4G	Red & Blue	Red, Blue, Green	3 cycles
4B	Red & Blue	Red, Blue, Green	10 cycles
5G	Magenta	Magenta, Green	4 cycles
5B	Magenta	Magenta, Green	12 cycles
8G	Red & Blue	Red, Blue	4 cycles
8B	Red & Blue	Red, Blue, Green	1 cycle
9G	Magenta	Magenta	4 cycles
9B	Magenta	Magenta, Green	10 cycles
10G	Magenta	Magenta, Green	3 cycles
10B	Red & Blue	Red, Blue, Green	12 cycles

The key distinction between the good and bad parameter sets was the persistence setting. The optimized values were to persist the contact for either three or four cycles, whereas the non-optimized values were one, ten, or twelve cycles. With three or four cycles of persistence, the target contacts formed a train of three or four dots, and as a new dot was plotted, a previous plot from three or four cycles ago was removed. With just one cycle of persistence, each new plot of a target contact was coincident with the removal of the previous plot, thus no train of dots ever formed. With ten or twelve cycles, target contacts remained on the display longer and the resulting train was ten or twelve dots in length.

During the experiment, the captured screen video was played back on a laboratory computer under experimenter control. The display resolution for the captured video was 760 x 760, both when captured and when played back.

2.3 Procedure

Each participant reviewed and signed the informed consent form prior to participation. The experimenter provided a brief overview and explanation of the experiment, and demonstrated the software, pointing out the appearance of a target-like object hopping across the screen. Participants were instructed to verbally report they spotted a moving object and to point to it on the display. Training was provided, initially featuring targets whose motion was relatively straightforward to spot. A second training

trial was provided, featuring target motion that was more complex. Participants were allowed to ask questions until they were confident they understood the task.

Each participant then viewed ten evaluation trials, consisting of five data sets presented twice each (once with good parameters and once with bad, as summarized in Table 1). The presentation order was randomized for each participant, constrained to ensure that each participant received a unique presentation order, and that presentation of a given data set with good versus bad parameters first was balanced across participants.

During a given trial, the experimenter confirmed that the participant was ready, and then initiated the video playback. The participant watched the video and verbally announced when he or she spotted a moving object. The experimenter recorded responses on a coded sheet that depicted the actual target location and movement for that trial. If the participant did not detect the object during the first replay (i.e., the first three loops), the presentation was repeated for an additional iteration of three loops. For each trial, the experimenter recorded which of the targets were detected (hits), which were missed, and any false alarms reported by the participants. The experimenter also noted if a correct detection was made immediately at the onset of data playback (i.e., before the playback restarted for the second loop.)

3 Results and Discussion

Performance across the five good and five bad trials is summarized in Table 2. Participants readily perceived the target in motion on nearly 95% of the good trials, and on less than 5% of the bad trials. In fact, only one of the ten participants correctly detected target motion on any bad trial. All but one participant correctly identified at least one target on all good trials; the one exception missed detecting a lone target on just one trial. All the correct detections occurred on the first iteration of the video playback.

Table 2. Summary of overall performance across evaluation trials

Trial ID	Positive ID	False Alarm	Miss	Immediate
4G	90%	0	10%	70%
5G	100%	3	0%	30%
8G	90%	0	10%	70%
9G	95%	0	5%	60%
10G	100%	1	0%	30%
4B	10%	1	90%	0%
5B	10%	0	90%	0%
8B	0%	0	100%	0%
9B	5%	3	95%	0%
10B	0%	0	100%	0%

There were only eight false alarms across all participants. Seven of the ten participants made no false alarm reports. One participant reported four of the eight false alarms.

As mentioned, the experimenter also recorded if a correct detection was reported immediately, that is, during the initial playback rather than after it had begun the second loop of playback. Over half the correct detections were made immediately. This finding indicates the participants intuitively perceived target identity and motion.

Across the range of non-optimized parameters that were tested, it was virtually impossible for participants to detect the target. Within the range of optimized parameters, though, the target became quite conspicuous – virtually impossible to miss, even for naïve observers. The persistence setting was the key parameter. With persistence set to three or four iterations, the resulting train of dots appears to hop across the screen, and preserved its identity as a moving object even when it changed direction or its color. With persistence set to just one cycle, each target contact plot disappeared as its successor appeared, and thus no train of dots was ever formed. For the longer persistence settings (ten or twelve cycles), a train of dots would form, but the high number of false contacts that were also appearing in close proximity interfered with object perception. With the optimized parameters, the target was perceived as a train of dots marching across the display. Participants readily perceived it as a single object that was moving across the display – not as a pattern of successive dots from which the presence of a target could be inferred. These perceptual objects emerged purely from the dynamic properties of the succession of dots, and were not dependent on color or any other attribute to reinforce the object permanence.

The playback speed was held constant in this experiment. In engineering development, it was clear that the combination of playback speed and persistence was a key combination. In particular, much slower playback speeds made it more difficult to perceive target motion. The playback speed chosen for the current experiment produced a target update rate of about 1.5 Hz. Our informal observations were that update rates of about 1.0 Hz still produced the effect reliably; below about 0.5 Hz, though, the effect was unreliable, even with optimized persistence settings. We plan to systematically vary this parameter in future experiments to get a better estimate of the range of effective combinations of update rates and persistence settings.

We also did not systematically vary the distance between the successive contact reports. We did allow targets to momentarily stop and then resume motion, but in general the moving targets moved at about the same speed while in motion. This parameter – the distance between the successive dots – is undoubtedly also important, and will need to be parsed in future experiments.

4 Conclusions and Recommendations

We were able to create dynamic perceptual objects by controlling the parameters of data presentation to produce an update rate of about 1.5 Hz for contacts of interest. This rate of presentation did not produce dynamic perceptual objects unless the persistence was set to preserve only two or three previous contacts. The result was the

perception of an object, consisting of three or four dots, which appeared to hop across the screen, perhaps changing color or direction. The apparent motion produced by this technique is likely related to beta motion, but the perception was of discrete jumps, not smooth motion. The perception was immediate and intuitive. This phenomenon can be used to help design displays that replay data with a large number of distractors in such a way as to allow the observer to readily detect a moving object in the presence of those distractors.

An obvious application is in presentation of data sets that involve geographic-based plots of potential contacts of interest over time. These data sets could be related to security threats, hard-to-detect contacts such as members of endangered species, or any other object of interest that is at least occasionally in motion and might be detected at a specific location at a specific time.

There is no particular reason why the x-y plane used for this technique must be geographic-based. It could be used for any coordinate system where systematic trends over time are of interest. For example, it could be applied to changes in economic data over time, or changes in opinion over time. For coordinate systems that are not geographic-based, the chosen axes should be continuous (for all practical purposes), and have meaningful ordinal properties.

Color can be used to help reinforce the identity of an object of interest, but color changes do not necessarily disrupt object perception. It is likely beneficial to use color to code attributes that change less frequently than the x-y coordinates used to plot the data – although we have not studied that assertion.

The focus of the display technique is to control data presentation rate so that an object of interest will change states at a rate in the range of 1.0 – 2.0 Hz. The presentation rate must be linked to a persistence parameter that promotes removal of stale data after three or four samples. When data already exist and are being presented, the appropriate playback speed and persistence setting may be calculable (or at least estimable) within the presentation software, by analyzing near neighbor state changes in the data set over time. When this technique is used to present real-time data, such calculations would necessarily be restricted to recent near neighbors.

The primary value of dynamic perceptual objects is that they make use of preattentive processing, that is, the observer does not have to concentrate and make deliberate inferences about the relationships depicted in the presentation. The natural, intuitive perception of motion allows cognitive resources to be devoted to other issues, such as the significance or impact of the depicted relationships. Dynamic perceptual objects can help observers quickly perceive attributes of interest in large data sets, and therefore can be a powerful visualization technique when presenting those data sets to interested observers.

References

1. Gestalt Psychology, http://en.wikipedia.org/wiki/Gestalt_psychology (retrieved February 7, 2014)
2. Kellman, P.J., Spelke, E.S.: Perception of Partly Occluded Objects in Infancy. Cognitive Psychology 15(4), 483–524 (1983)

3. Moscovitch, M., Wincour, G., Behrmann, M.: What is Special about Face Recognition? Nineteen Experiments on a Person with Visual Object Agnosia and Dyslexia but Normal Face Recognition. Journal of Cognitive Neuroscience 9(5), 555–604 (1997)
4. Carswell, M.C., Wickens, C.D.: The Perceptual Interaction of Graphical Attributes: Configurality, Stimulus Homogeneity, and Object Integration. Perception & Psychophysics 47(2), 157–168 (1990)
5. Eden, B.: Chapter 1: Information Visualization. Library Technology Reports 41(1), 7–17 (2005)
6. Robertson, G., Czerwinski, M., Fisher, D., Lee, B.: Selected Human Factors Issues in Information Visualization. Reviews of Human Factors and Ergonomics 5(1), 41–81 (2009)
7. Phi Phenomenon, http://en.wikipedia.org/wiki/Phi_phenomenon (retrieved February 7, 2014)
8. Beta Movement, http://en.wikipedia.org/wiki/Beta_movement (retrieved February 7, 2014)

Different Roles of Foveal and Extrafoveal Vision in Ensemble Representation for Facial Expressions

Luyan Ji[1,2], Wenfeng Chen[1,*], and Xiaolan Fu[1]

[1]State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
{jily,chenwf,fuxl}@psych.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. People could extract mean expression of multiple faces pretty precisely. However, the mechanism of how we make such ensemble representation was far from clear. This study aimed to explore how faces in the foveal and extrafoveal vision contribute to the ensemble representation and whether the emotion of faces modulates the contribution. In the experiment, the expressions of foveal and extrafoveal faces were independently manipulated by changing the ratio of happy vs. angry faces. The participants reported whether the overall emotion was positive or negative. The results showed that faces in the foveal vision were given more weight than those in the extrafoveal vision in ensemble emotional representation. In addition, the ensemble perception was more accurate when faces in the extrafoveal vision were positive. These findings have great implications for the emotional design in interactive systems, especially when there are multiple users or multiple avatars presented on the screen.

Keywords: Ensemble representation, Facial expression, Foveal vision, Extrafoveal vision.

1 Introduction

Emotion is very important in our interaction with people as well as computers in everyday life. Recognizing users' emotions is the first step towards using emotions to improve adaptive performance of computers. Usually, the research on emotion recognition has been focused on a single person or an isolated face. However, in some occasions, like group studies or multiplayer games, there might be a number of people interacting with computer system at the same time. Nevertheless, the mechanism of how we acquire such ensemble representation was far from clear. For example, not all faces could fall into the foveal vision at a time, but little work concentrated on the role of locations in the visual field in ensemble perception. In this study, we aimed to investigate what the roles foveal faces and the extrafoveal faces play respectively in ensemble representation for facial expressions. It has great implications for the

* Corresponding author.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 164–173, 2014.

emotional design in interactive systems, especially when there are multiple users or multiple avatars presented on the screen.

Previous studies have demonstrated that observers can extract the mean emotion of multiple faces rather precisely [1-3]. This ability is robust and flexible, operating even on sets containing as many as 24 faces shown for only 100 ms [4]. However, when the set size is relatively large, observers cannot focus on all the faces simultaneously shown at one time. During normal viewing, our eyes direct a sequence of images to the fovea when sampling the visual scene and the fovea is a small region of the retina that corresponds to the central 2° of the visual field [5]. When the exposure time of face set was sufficient (e.g., a set of four faces shown for 2000 ms [2]), the observers could have time to freely scan every face; but if the set made up of large number of faces was shown very briefly [4], the observers could only glimpse a few faces or even some parts of a face. In the latter condition, when some faces or some features of one face occupy the fovea, on one fixation, other faces in the set would project extrafoveally.

The visual system indeed has the ability to detect or recognize emotional facial expressions at extrafoveal visual field, but the performance declined with increased eccentricity [6-8]. Neuroimaging study have shown that the affective modulation of early ERP components exists for both centrally and peripherally presented pictures, but the latter was lower in amplitude and slightly delayed [7]. The declined performance and neural response may be due to low visual acuity and low contrast sensitivity of extrafoveal vision [9]. The parvocelullar system which mainly begins in central part of the retina is particularly sensitive to contrast and high spatial frequency, like the detail of objects [10]; on the contrary, the magnocellular system which essentially originates from peripheral parts of the retina is less effective in contrast detection, but possesses a high temporal resolution [11-12]. According to these facts, it can be speculated that faces would be processed differently in the extrafoveal vision, contrast to in the foveal vision.

However, little is known about whether the foveated faces and the nonfoveated faces would contribute differently to ensemble representation for facial expressions. To our best knowledge, only one study paid attention to the relationship between eccentricity and the averaging of emotional expressions [13]. They asked the observers to judge the emotional intensity of the target face, which was shown isolated or surrounded by flankers, and the face (set) was presented either centrally or extrafoveally. The results revealed that judgment of the target face was less accurate and leaned more towards the average emotion of the flanker set when presented in the extrafoveal location. It seems to suggest that the impact of extrafoveal ensemble faces on individual recognition was greater than that of central ensemble faces. But it remains unclear as for the roles of the extrafoveal faces and foveal faces in ensemble representation. To tackle this issue, participants were instructed to report the ensemble emotion of a face set in the present study, instead of judging on one target face. We also developed a new display that both foveal and extrafoveal vision can be used, and that was different from gaze-contingent display [14-15], which restricted the observers to use central vision only or peripheral vision only. In addition, the central four faces would always occupy the foveal visual field by carefully manipulating the visual angle, and the remained surrounding faces would fall out of the fovea on one

fixation. With this kind of display, the study could explore how observers represented the mean emotion when some faces were in foveal location while the other faces were in extrafoveal location.

Another purpose of this study was to investigate whether the role of the extrafoveal faces in the ensemble emotional representation would be enhanced or weakened by the emotional valence. It was evident that recognizing negative expressions (like anger, fear and sadness) is impaired when presented peripherally, whereas recognition of happiness suffered the least from peripheral presentation and was comparable with foveal performance [16]. However, controversy remains whether the extrafoveal vision favors for happy faces or for the negative faces. When an emotional face paired with a neutral face was shown in bilateral extrafoveal vision, happy face was identified faster than negative face indicated by shorter saccade latencies [17]; but the study adopting attention probe technique showed an advantage in rapid orienting of attention towards negative face, but not positive face [18]. There is also evidence suggesting that positive and negative faces showed similar efficiency in attentional capture, although both were more effective than neutral faces [19]. In general, it is far to reach a consistent conclusion on the role of extrafoveal facial emotions. Furthermore, few studies have focused on the attentional bias of ensemble representation for multiple facial expressions. So far, only one study found a positive bias of ensemble emotional representation, but did not explore whether ensemble emotional representation suffered from extrafoveal vision [4]. Therefore, the relationship between ensemble representation and extrafoveal emotion requires further investigation.

In the present study, we used happy and angry faces as emotion stimuli to investigate whether the foveal faces and the extrafoveal faces contribute differently in ensemble representation for multiple emotional expressions and whether there was an advantage for extrafoveal happy face or angry face in ensemble coding. Similar as Yang et al [4], we used face images from different people rather than morphed images of one person. We hypothesized that faces in the foveal vision, compared with faces falling out of the fovea, would be given higher weight in ensemble representation for facial expressions. Therefore, when the emotion of foveal faces was incongruent with that of extrafoveal faces, observers would judge the overall emotion based more on foveal faces. For the second question, there might be several possibilities: (a) The role of extrafoveal happy faces rather than angry faces was less impaired in the averaging of facial expressions, because the happy faces in the extrafoveal vision enjoyed an advantage in identification [16-17], (b) the role of extrafoveal angry faces rather than happy faces was less impaired in the ensemble representation for facial expressions, because the angry faces captured attention better than happy faces in the extrafoveal vision [18] , and (c) the role of extrafoveal happy and angry faces was impaired to the same extent, because the advantages were for both sides were comparable [19].

2 Method

2.1 Participants

Forty-six volunteers (19-26 years old; 23 females) participated in the experiment. All of them had normal or corrected-to-normal vision and were paid for their time. Four

participants did not properly respond to the face set (two responded too fast, and two too slow, which exceeded three standard deviations). Final analyses were based on the remaining 42 participants.

2.2 Stimuli

We selected 32 color photographs of faces from BU3DFE database [20], including 8 happy male faces, 8 happy female faces, 8 angry male faces and 8 angry female faces. All images were scaled to the same mean luminance and root-mean-square contrast [21]. Each face image subtended a visual angle of 1.97° × 1.97° at a viewing distance 85cm, and was presented against a black background.

Faces were presented in a 4 × 4 invisible grid, and the set of 16 face images occupied 6.40° × 6.53° of visual angle. For the central four faces, they occupied 3.98° × 4.02° of visual angle, falling into the foveal visual field (1.5mm diameter, about 4.3 degree [22]) and the eccentricity of each was 1.44° from the fixation point (Figure 1). The mean emotion of central four faces and that of surrounding twelve faces was independently manipulated by changing the ratio of happy to angry faces (Table 1). The ratio of happy vs. angry faces in the whole set was 3:1, 5:3, 3:5, or 1:3. Due to the number of surrounding faces outweighed that of central faces, the overall emotion of the whole set was always consistent with that of surrounding faces.

Face images in each set were randomly selected with two constraints: (1) an equal number of male and female models were presented in each set, and (2) no two images in each set were of the same model.

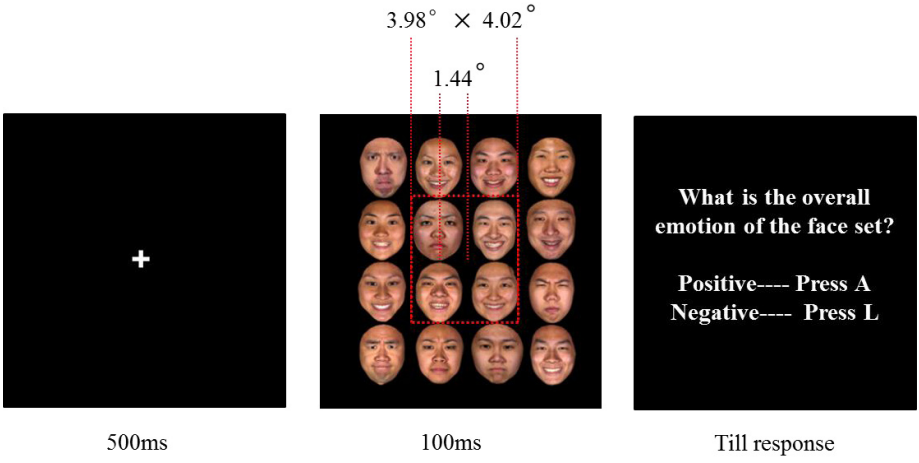


Fig. 1. Experiment procedure and stimuli. In the sample face set, the mean emotion of central four faces is positive and that of surrounding faces is also positive. It is a congruent condition.

2.3 Design

The design involved two within-subject factors: congruency (congruent vs. incongruent) and extrafoveal emotion (positive vs. negative).

Table 1. Ratio of happy vs. angry faces in the foveal and extrafoveal vision respectively and the corresponding mean emotion of foveal and extrafoveal faces

Foveal H vs. A	Extrafoveal H vs. A	Foveal emotion	Extrafoveal emotion	Congruency[1]
4:0	8:4	Positive	Positive	C
4:0	2:10	Positive	Negative	IC
4:0	0:12	Positive	Negative	IC
3:1	9:3	Positive	Positive	C
3:1	7:5	Positive	Positive	C
3:1	3:9	Positive	Negative	IC
3:1	1:11	Positive	Negative	IC
1:3	11:1	Negative	Positive	IC
1:3	9:3	Negative	Positive	IC
1:3	5:7	Negative	Negative	C
1:3	3:9	Negative	Negative	C
0:4	12:0	Negative	Positive	IC
0:4	10:2	Negative	Positive	IC
0:4	4:8	Negative	Negative	C

2.4 Procedure

Each trial began with a 500ms fixation point at the center of the screen, followed by a set of faces presented for 100ms. Due to the short duration, participants were unlikely to execute a saccade (minimal saccade latency is about 150 ms [23]) to make the

[1] C denotes congruency and IC denotes incongruency. When the mean emotion of the foveal and extrafoveal faces are both positive or negative, it is the congruent condition. When the mean emotion of the foveal is positive whereas that of extrafoveal faces is negative, or vice versa, it is the incongruent condition.

surrounding 12 faces in the foveal vision. Immediately after the face set disappeared, participants were prompted to indicate whether the overall emotion of previous face set was positive or negative by pressing corresponding keys ("A" for positive and "L" for negative, vice versa).

Participants completed 280 trials, with 120 trials evenly divided into two congruent conditions and 160 trials evenly divided into incongruent conditions. There were 16 practice trials before the main experiment.

3 Results

3.1 Accuracy

The accuracies were calculated for each subject and each condition based on whether the subjective report conforms to the mean emotion of the whole set. A repeated-measure of ANOVA (congruency and extrafoveal emotion) was conducted. The results showed that the accuracy was significantly higher when the foveal and extrafoveal emotion were congruent than incongruent (F (1, 41) = 162.54, p<.001, η_p^2 = .80), and the accuracy was significantly higher when extrafoveal faces were happy than those were angry (F (1, 41) = 9.79, p<.01, η_p^2 = .19); but there was no interaction of congruency by extrafoveal emotion (F (1, 41) <1) (Figure 2). A t-test revealed that the accuracy was significantly lower than the chance level in the incongruent conditions (t (41) = -7.49, p< .001; t (41) = -5.92, p< .001).

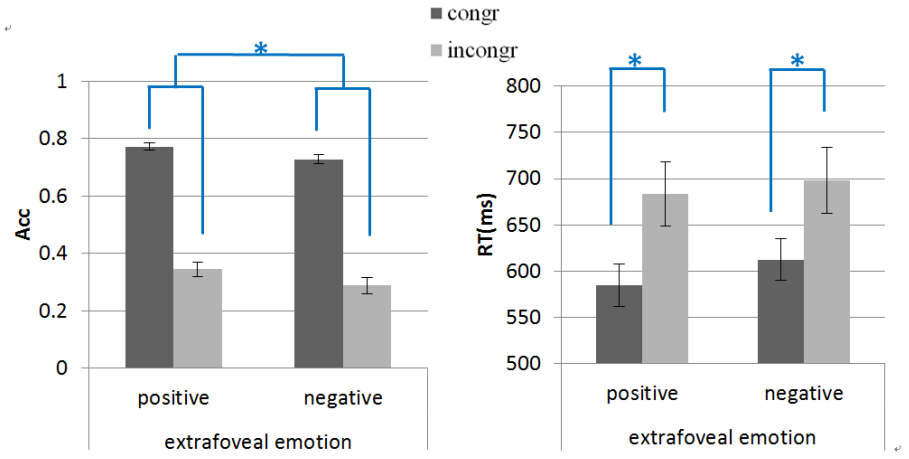


Fig. 2. Mean accuracies and RTs as a function of congruency and extrafoveal emotion. Error bars indicate ±1 standard error. Star marks indicate significant results.

3.2 Reaction Times

The reaction times for correct responses for each condition are shown in Figure 2. We conducted a 2 (congruency) by 2 (extrafoveal emotion) analysis of variance

(ANOVA) for repeated measures. The analysis revealed that the reaction times were faster when the foveal and extrafoveal emotion were congruent than incongruent (F (1, 41) = 24.70, p<.001, η^2 = .38); but there was no difference between extrafoveal happy face and extrafoveal angry face condition in reaction times (F (1, 41) = 3.10, p=.086, η^2 = .07) nor interaction between congruency by extrafoveal emotion (F (1, 41) <1).

4 Discussion

The present study showed that the foveal faces were given more weights than extrafoveal faces in ensemble representation. When the emotion of foveal faces and extrafoveal faces were incongruent, emotion in the foveal vision contributed more to the averaging, although the overall emotion was consistent with extrafoveal emotion rather than foveal emotion, so the accuracy was much lower and reaction times were much slower in incongruent conditions compared with congruent conditions. Besides, compared to extrafoveal angry faces, we were more sensitive to extrafoveal happy faces, and the happy expressions enhanced the role that extrafoveal faces played in ensemble representation.

4.1 Foveal Faces Contribute More to Averaging Facial Expressions

We found a below chance performance of ensemble representation when the foveal emotion and the extrafoveal emotion were incongruent, and indicated that subjects did not acquire a correct ensemble emotional representation of the face set. Since the overall emotion was always consistent with the extrafoveal emotion, the wrong response may suggest that subjects were prone to judge the overall emotion based on the foveal emotion, which was opposite to the extrafoveal emotion.

Previous research has found that not all items in a set contribute the same to the mean. The mean size estimations would bias towards the size of the attended item [24], and the emotional outliers in a set of faces would be discounted in the summary representation [25]. Besides, a recent eye movement study showed that the perceived ensemble expression is based on the particular faces that are fixated when the faces are randomly arranged [26]. In the present study, we are the first to explore the role of foveal and extrafoveal vision in ensemble representation for facial expressions. Consistent with hypothesis, the foveal faces played a much more important role in the ensemble representation compared with extrafoveal faces. The central four faces got the fixation and enjoyed the foveal vision; while because of limited duration time, the surrounding faces could not be fixed and were only exposed extrafoveally. The magnocellular system which essentially originates from peripheral parts of the retina was low in sensitivity to high spatial frequency [12], and removal of high spatial frequency would lead to great impairment in discriminating and recognizing facial expressions [16].

The faces in the extrafoveal vision are not only detected or recognized worse [6, 8], but are also down weighted in the ensemble representation, even though the number of extrafoveal faces are much larger. This result also suggests that the statistical properties of a set of items are not always extracted automatically, since the mean emotion of multiple faces should not be affected by foveal or extrafoveal vision. But there is a

potential alternative that it is the difference of difficulty in foveal and extrafoveal processing that counts the different role of foveal and extrafoveal in the ensemble represenation, if the duration time was extended a little (but still within the minimal saccade latency) or the number of extrafoveal faces were reduced, observers might extract the mean emotion in a parallel way with distributed attention and then the different roles of foveal and extrafoveal vision might disappear. This alternative remains to be investigated in further research.

4.2 Advantage of Extrafoveal Happy Faces

Although the role of extrafoveal faces in the ensemble representation was impaired, the present study demonstrated that the role of extrafoveal happy faces was less impaired, compared with extrafoveal angry faces, no matter the congruency. So, the ensemble representation does not shut the door completely to the extrafoveal emotion.

The advantage for extrafoveal happy faces might due to physical feature saliency as well as positive affect. All the happy face images we used in the present study had a smile with an open mouth. It might be not necessary to recognize a happy face, but just detect the salient smile [16]. Especially when multiple happy faces were shown together, the open mouth with white teeth would be rather salient. On the other hand, some researchers pointed out that apart from featural processing, the affective processing plays a role that the priming effect of extrafoveal happy faces emerged at 750ms rather than 250ms [14]. In the present study, we are not aimed to explore whether the featural or affective processing contribute more to the advantage for extrafoveal happy faces, but it is worth further investigation and especially when there are multiple extrafoveal happy faces.

Then, why did not the angry faces enjoy an advantage in the extrafoveal vision? There are several possibilities. Firstly, the participants in the present study were generally in low level of anxiety and the angry faces might be only minor threatening. Different from the vigilance for severe threat [27], low vulnerable groups are prone to avoid the minor threat [28], so the processing of extrafoveal angry faces might be inhibited rather than strengthened. Secondly, the angry faces in the extrafoveal vision were not that salient compared with extrafoveal happy faces, and the physical saliency competed against the affective significance. Last but not least, extrafoveal angry faces hold attention and lead to a "disengagement deficit" from such stimuli [29], so processing extrafoveal angry faces might come at a cost. Future work will be needed to clear these possibilities and found out why there was an advantage for extrafoveal happy faces compared with angry faces. Maybe, we could use more threatening stimuli, like fearful faces, as well as anxious participants, to explore whether the advantage for extrafoveal happy faces still exist or whether there would be an advantage for extrafoveal negative faces instead.

4.3 Implications of the Present Study

The study found that foveal faces contribute more to the ensemble representation for facial expressions, and the extrafoveal happy faces were less impaired compared with extrafoveal angry faces. It also gives a new insight to the emotional design in interactive systems. For multiple users' interaction with computer, the computer has

to combine the overall emotional information from all these users, but may give different weights to each individual. Users who stay closer to the screen may be more involved in interacting with the computer, so it makes sense for the computer to give more weights to the emotion expressed by these users when analyzing the overall emotional state of the users.

In addition, if there are multiple avatars showed on one screen, the designers should taker users' attentional capacity and visual field into consideration. Our visual system has limited attentional and short-term memory capacity [30-31]. If there is too much information on the computer screen, our brain may take use of the statistical regularity to condense the redundant information [32], but the extrafoveal information will be down-weighted or ignored to a large extent. So, the most important information or the key avatars should be arranged in the foveal or near the foveal vision of users.

On the other hand, the happy faces or other positive information shown extrafoveally play a more important role compared with extrafoveal negative information. When receiving a negative feedback from the computer, the users may feel better if there was also something positive, even if not presented at the corner of the screen.

Acknowledgement. This research was supported in part by grants from National Basic Research Program (2011CB302201), and the National Natural Science Foundation of China (31371031).

References

1. Haberman, J., Harp, T., Whitney, D.: Averaging Facial Expression over Time. Journal of Vision 9(11), 1–13 (2009)
2. Haberman, J., Whitney, D.: Rapid Extraction of Mean Emotion and Gender from Sets of Faces. Current Biology 17(17), R751–R753 (2007)
3. Haberman, J., Whitney, D.: Seeing the Mean: Ensemble Coding for Sets of Faces. Journal of Experimental Psychology: Human Perception and Performance 35(3), 718–734 (2009)
4. Yang, J.W., Yoon, K.L., Chong, S.C., Oh, K.J.: Accurate but Pathological: Social Anxiety and Ensemble Coding of Emotion. Cognitive Therapy and Research 37, 572–578 (2013)
5. Atkinson, A.P., Smithson, H.E.: Distinct Contributions to Facial Emotion Perception of Foveated versus Nonfoveated Facial Features. Emotion Review 5(1), 30–35 (2013)
6. Bayle, D.J., Schoendorff, B., Henaff, M.A., Krolak-Salmon, P.: Emotion Facial Detection in the Peripheral Visual Field. PLoS ONE 6(6), e21584 (2011)
7. Rigulot, S., Delplanque, S., Despretz, P., Defoort-Dhellemmes, S., Honore, J., Sequeira, H.: A Spatiotemporal Analysis of Early ERP Responses. Brain Topography 20, 216–223 (2008)
8. Rigulot, S., D'Hondt, F., Honore, J., Sequeira, H.: Implicit Emotional Processing in Peripheral Vision: Behavioral and Neural Evidence. Neuropsychologia 50, 2887–2896 (2012)
9. Robson, J.G., Graham, N.: Probability Summation and Regional Variation in Contrast Sensitivity across the Visual Field. Vision Research 21, 409–418 (1981)
10. Dacey, D.M., Petersen, M.R.: Dendritic Field Size and Morphology of Midget and Parasol Ganglion Cells of the Human Retina. Proceedings of the National Academy of Sciences 89, 9666–9670 (1992)
11. Livingston, M.S., Hubel, D.H.: Psychophysical Evidence for Separate Channels for the Prception of Form, Color, Movement, and Depth. Journal of Neuroscience 7, 3416–3468 (1987)

12. Schiller, P.H., Logothetis, N.K., Charles, E.R.: Functions of The Colour-opponent and Broad-band Channels of the Visual System. Nature 343, 68–70 (1990)
13. Fielding, K.M., Carvey, R.J., Liu, C.H.: Singular or Summary: Averaging of Facial Expression in Sets is Modulated by Eccentricity. Journal of Vision 13(9), 588 (2013)
14. Calvo, M.G., Nummenmaa, L., Avero, P.: Recognition Advantage of Happy Faces in Extrafoveal Vision: Featural and Affective Processing. Visual Cognition 18, 1274–1297 (2010)
15. Reingold, E.M., Loschky, L.C., McConkie, G.W., Stampe, D.M.: Gaze-contingent Multiresolutional Displays: An Integrative Review. Human Factors 45, 307–328 (2003)
16. Goren, D., Wilson, H.R.: Quantifying Facial Expression Recognition across Viewing Conditions. Vision Research 46, 1253–1262 (2006)
17. Calvo, M.G., Nummenmaa, L.: Eye-movement Assessment of the Time Course in Facial Expression Recognition: Neurophysiological Implications. Cognitive, Affective & Behavioral Neuroscience 9(4), 398–411 (2009)
18. Pourtois, G., Grandjean, D., Sander, D., Vuilleumier, P.: Electrophysiological Correlates of Rapid Spatial Orienting towards Fearful Faces. Cerebral Cortex 14, 619–633 (2004)
19. Vuilleumier, P., Schwartz, S.: Emotional Facial Expressions Capture Attention. Neurology 56, 153–158 (2001)
20. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial Expression Database for Facial Behavior Research. In: 7th IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2006), pp. 211–216. IEEE Press, Southampton (2006)
21. Bex, P.J., Makous, W.: Spatial Frequency, Phase, and the Contrast of Natural Images. Journal of the Optical Society of America 19, 1096–1106 (2002)
22. Riordan-Eva, P.: Anatomy & Embryology of the Eye. In: Riordan-Eva, P., Cunningham, E.T. (eds.) Vaughan & Asbury's General Ophthalmology, 18th edn., p. 13. McGraw-Hill Medical (2011)
23. Rayner, K.: Eye Movements in Reading and Information Processing: 20Years of Research. Psychological Bulletin 124, 372–422 (1998)
24. de Fockert, J.W., Marchant, A.P.: Attention Modulates Set Representation by Statistical Properties. Perception & Psychophysics 70(5), 789–794 (2008)
25. Haberman, J., Whitney, D.: The Visual System Discounts Emotional Deviants When Extracting Average Expression. Attention, Perception, & Psychophysics 72(7), 1825–1838 (2010)
26. Wolfe, B., Kosovicheva, A.A., Leib, A.Y., Whitney, D.: Beyond Fixation: Ensemble Coding and Eye Movements. Journal of Vision 13(9), 710 (2013)
27. Koster, E.H.W., Crombez, G., Verschuere, B., DeHouwer, J.: Selective Attention to Threat in the Dot Probe Paradigm: Differentiating Vigilance and Difficulty to Disengage. Behaviour Research and Therapy 42(10), 1183–1192 (2004)
28. Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M.J., van IJzendoorn, M.H.: Threat-related Attentional Bias in Anxious and Nonanxious Individuals: A Meta-Analytic Study. Psychological Bulletin 133(1), 1–24 (2004)
29. Fox, E., Russo, R., Bowles, R., Dutton, K.: Do Threatening Stimuli Draw or Hold Visual Attention in Subclinical Anxiety? Journal of Experimental Psychology: General 130(4), 681 (2001)
30. Luck, S.J., Vogel, E.K.: The Capacity of Visual Working Memory for Features and Conjunctions. Nature 390(6657), 279–281 (1997)
31. Scholl, B.J., Pylyshyn, Z.W.: Tracking Multiple Items through Occlusion: Clues to Visual Objecthood. Cognitive Psychology 38(2), 259–290 (1999)
32. Haberman, J., Whitney, D.: Ensemble Perception: Summarizing the Scene and Broadening the Limits of Visual Processing. Chapter to Appear in an Edited Volume. A Festschrift in Honor of Anne Treisman (2011)

The Effect of Driving Speed on Driver's Visual Attention: Experimental Investigation

Doori Jo[1], Sukhan Lee[2,*], and Yubu Lee[3]

[1] Department of Interaction Science, Sungkyunkwan University, Seoul, Republic of Korea
jdl6427@gmail.com
[2] College of Information and Communication Engineering,
Sungkyunkwan University, Suwon, Republic of Korea
lsh1@skku.edu
[3] Center for Neuroscience Imaging Research, Institute for Basic Science,
Sungkyunkwan University, Suwon, Republic of Korea
basilia@skku.edu

Abstract. It has been reported that the increase in driving speed incurs a shortened pupil distance, termed as a visual tunneling phenomenon. However, our experimental investigation shows that the effect of driving speed on driver's visual attention should be understood in terms of the maximum field of view that can balance against the maximum amount of visual information a driver can take/handle against. More specifically, our experimentation shows the following: For the sake of ensuring safety, drivers tend naturally to take as much visual information as possible, should it be allowed in terms of the maximum amount of visual information they can take/handle. However, the maximum visual information a driver can take/handle is different among individuals according to their level of driving expertise. Since the increase of driving speed increases the amount of visual information to process, a driver may be able to expand their field of view only up to the point where the amount of visual information to process balances the maximum amount he/she can take/handle. Beyond this point, the increased anxiety stress may even further reduce the maximum visual information a driver can take/handle, thus further diminishing the field of view, leading to a tunneling effect.

Keywords: Visual attention, Visual information processing, Mental workload, Driving expertise.

1 Introduction

Sometimes drivers fail to allocate their attention optimally. Because of the driver's knowledge of the current road and traffic conditions affects driver's attention endogenously and sudden changes in the visual field cause exogenous shift of attention. These two categorized factors influence change of the driver's attention allocation.

* Corresponding author.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 174–182, 2014.

The research on visual interference caused by endogenous factors provides a more direct approach to study higher level interference processes. If the distractors have no explicit foveal load, then their effects are hard to be explained. In addition to this theoretical implication, being engaged in one's own affairs, concerns and caused emotions while driving has, in itself, an applied interest because it represents a common everyday situation that has received little attention in the research field [1].

Using concurrent tasks with no foveal load with mental workload, some studies directly approach the relationships between attention and eye movement including gaze in real driving [2]. It also studied that the effects of different tasks causing mental workload on visual behavior and driving performance [3]. Following previous studies, several measures of visual search behavior were affected by workload caused by mental task and one of the general effects was a spatial gaze concentration like visual tunneling. Visual tunneling can be considered a plausible mechanism to optimize visual resource allocation by increasing the priority assigned to the road ahead. In contrast, the eventual negative value of reduced inspection of peripheral areas should also be considered. Likewise, driver's mental workload can affect to their attention and cognitive resource. Many current studies on a driver's mental workload also reported that mental workload has been shown to play a substantial role in driving safety [4]. So, It is very important that understanding mental workload's trait and relationship between mental workload and driver's attention.

Normally, when the mental workload is high, cognitive resource is reduced. Reduced cognitive resources by high workload would influence drivers' anticipations of emergent problems and their use of knowledge to avoid hazards Once the mental workload reaches an unacceptable level, driving safety may suffer [4,5,6]. A driver's mental workload can be influenced by many factors such as complex driving environment, talking on the phone and emotional states.

Following the Eysenck and Calvo's Processing Efficiency theory, there is explanation for how driver's emotion and emotionally involved behavior could negatively affect driving [7]. It suggests that when an individual experiences anxiety or stress they are less efficient in processing incoming sensory information and have to work harder to maintain performance levels. They claim that this is because anxiety leads to a depletion of central executive resources: these resources are used to cope with the increase in 'Cognitive anxiety' that is experienced. As a result, the individual must share resources between tasks. While the experience of anxiety could lead to the individual consciously applying more effort to the task in hand, and thus 'reinvesting' in controlled processing [8], it could also lead to greater distraction from the task as Central executive resources that are directed towards the task of driving are depleted by the presence of anxiety, making the individual more prone to distraction and thus resulting in poorer performance [7]. The effects of increased anxiety on task goals and performance have been investigated, with findings suggesting that increased anxiety introduces task-irrelevant goals which compete with task-relevant goals, depleting Central executive resources [9]. This increase in overall workload, in turn, contributes to failures in spatial working memory [10] and decreased visual awareness [2; 11, 12].

Following the Easterbrook's study also, emotional arousal reduces the range of visual cues that are used by an individual when scanning a visual scene [13]. He

argued that in some cases this can be adaptive, as it would enable irrelevant visual information to be ignored. However in situations in which a range of visual cues are required for successful execution of the task, such as driving, this reduced scanning could have a detrimental effect on performance. Derryberry and Tucker's [14] model was suggests, when an individual experiences high-arousal negative emotion, sensory processing is reduced and fewer cognitive resources are made available for task completion [15]. Janelle, Singer, and Williams [16] support this view with findings from their simulator study. They found that high anxiety 'drivers' demonstrated visual tunneling (a narrowing of their visual attention, measured by tracking eye movements) but paradoxically also showed a greater tendency to be distracted by irrelevant cues in peripheral vision than did non-anxious participants. Anxious participants tended to focus on the central field (showing attentional narrowing). As a result, when any event occurred in the periphery, regardless of its level of relevance, they had to shift their gaze entirely to the peripheral field in order to process the information. Non-anxious participants did not demonstrate such a shift in gaze pattern. Janelle et al. claim that this shows that the experience of anxiety brings about hyperdistractibility [17]. Thus, evidence suggests that anxiety causes increasing of mental workload of individuals and decreasing task performance because of visual allocation's change such as visual tunneling in driving tasks.

Previous literature [18] suggested that driver's anxiety and stress level is affected by environment while driving such as driving speed. When drivers increased their speed, they tended to employ anxious behaviors such as check their mirrors more frequently. As anxiety consumes cognitive resources [18] and as safe driving requires sustained attention and emotional composure, it was hypothesized that higher levels of driving speed would be related more an increased driver's anxiety.

Until now, we have investigated that the driver's visual attention change influenced by mental workload from emotional state and relationship with driving speed and emotion. The present paper focuses upon mental workload caused by anxiety from the different speed levels and relationship between mental workload and driving experience. To investigate speed's influence on driver's visual attention, we measured driver's eye movement using eye tracker depending on speed levels under the target focusing task in real driving situation. And we also measured driver's anxiety level about driving speed through the questionnaire. It measured for relationship between driver's visual attention and anxiety caused by speed condition. Lastly, we also described the difference with experienced drivers and novice drivers shortly. More specific contents are in the next part.

2 Method

2.1 Participants

7 participants (3 male, 4 female) from the Cheon-an City of South Korea were recruited. 3 of them were novice driver who have 5 times driving experience in maximum and the other were experienced driver who have over 5 years driving experience and participants were naive to the purposes of the study. They received course credits

for participating. Participants ranged in age from 21 to 56 years (M = 35.86 years, SD = 14.40 years). All held a valid Republic of Korea driving license.

Table 1. Participants' Infomation

No.	Age	Gender	Driving experience
1	56	M	32 years
2	26	F	2 times
3	52	F	10 years
4	26	F	4 times
5	21	M	3 times
6	44	M	24 years
7	26	F	3 years

2.2 Design

This study used a mixed experimental design. There were two independent variables: Target type (typed by target distance, with two levels: window target, 20m distance target) and Speed condition (with three levels: 0km/h, 40km/h, 60km/h). Each participant completed three sessions in a real way driving situation.

The dependent variable was eye movements (pupil distance). During the target focusing, eye movement was monitored by eye tracker. This experiment designed for understanding of relationship between speed and mental workload caused by anxiety through the pupil distance measuring for visual tunneling.

2.3 Apparatus

Questionnaires. The State-Trait Anxiety Inventory − form Y1[19] :this is a 20-item questionnaire, using a 4 point likert scale, designed to measure the individual's current psychological state. This measure has also been validated and found to be reliable (reliability = .86). The questionnaire was given both before and after completion of the driving tasks to ascertain any changes in anxiety levels following the procedure.

Eye tracking equipment. All participants had their eye movements tracked using a SMART EYE PRO eye tracker (developed by SMART EYE Cop., Sweden). 3 Cameras with 2 IR flashes was attached on top of the dash board and recorded what participants saw from their own viewpoint and the eye movements of participants were sampled, from the both eyes, at a rate of 60 Hz. The temporal resolution of the eye movement equipment was 25 Hz, and the spatial accuracy was 1_. For the purposes of analysis, the position of both pupil centers was estimated and from that, pupil distance was calculated.

3 Procedure

Each participant completed the experiment individually, in a 30-min session. Participants first completed the Anxiety questionnaire about driving speed, in order to assess their anxiety state about speedy driving. Then, driving course and task was explained to the participant, and they were given a brief period to become accustomed to the eye tracking system. After that, experiment was begun. First, participants were informed that they should focus on the 2 type of target with 3 speed conditions. Before the driving, they stared two type of target (on the window, on 20m distance) few seconds and eye tracker recoded that information. Then, participant was informed to be focus on the each target for a few seconds when they drive on 40km/h and 60km/h. In the driving situation, distance target was on the car which in front of the experiment car. Targeted car was droved at the same speed with experiment car, so distance between two car was maintained. In this session also monitored and recorded by eye tracking system. When the records was finished, participants was checked their anxiety state by questionnaire then whole session was done.

4 Result

4.1 Anxiety Questionnaire

A 2x2 mixed ANOVA was carried out on the anxiety data with repeated measures on the scores from the STAI (before and after driving) to all participants. It was found that there was a significant difference in anxiety scores before and after completion of the experiment ($p < .01$) but there was no significant difference between experienced drivers and novice drivers ($p > .05$).

4.2 Pupil Distance

Average of the participants pupil distance was increased following the driving speed increasing in two target types both (figure 1). From this, there was no characteristic about visual tunneling phenomenon at the high speed condition and driver's mental workload was not influenced even that is high speed.

And when participants focused on the nearer target, pupil distance was shorter than when they focused on distance target. It is caused by human eye movement's trait called convergence accommodation which is pupil distance getting narrower when human focus on the nearer object. But, even concerning this natural trait, driver's eye movement was affected by driving speed. Under the eye convergence accommodation in near target type, pupil distance also increased by driving speed. We interpreted from this result, driver's unconscious behavior for keep their safety, it was explained in next part more specifically.

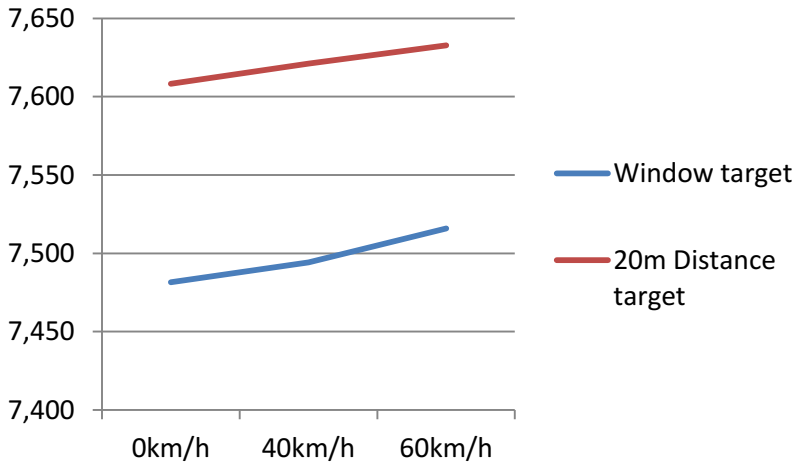


Fig. 1. Average of Subject's pupil distance

There is also considerable result in near target type. Comparing increasing degree of pupil distance near target type and distance target, near target type's increasing degree was increased rapidly at 60km/h. This eye movement was observed from most of the participant (Figure 2), it is unexpected result which is far from the hypotheses expected from previous studies.

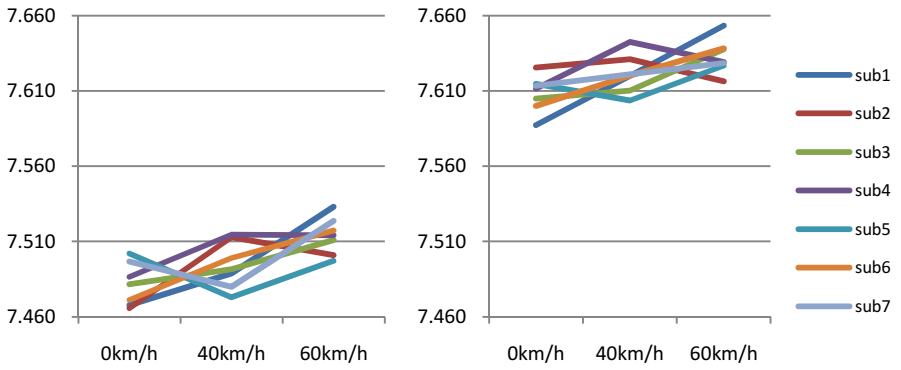


Fig. 2. Pupil distance of subjects (Left: Window target, Right: Distance target)

But, when we concern about each participant driving experience, result can be interpreted in interesting way. In Figure 2, every subject's pupil distance depending on target type and speed condition is appeared. We could get information about each subjects from the table 1. Following the table 1, subjects 2, 4, 5 are novice driver and they show different pupil distance change than experienced drivers. Especially subject

2 and subject 4 showed similar pupil distance change which can be interpreted visual tunneling phenomenon when the speed is getting higher in distance target type. But interestingly, it was not observed in near target type. From now on, result from the experiment was reported. More controvertible things will be deal with in next part.

5 Discussion

We expected that shortened pupil distance which concerned with visual tunneling phenomenon when the speed went higher than before. But from the result, we considered there was no remarkable influence (which represented by visual tunneling phenomenon) of the anxiety caused by speed in average. From this, here are two implications. First, adequately high driving speed is may not a considerable mental workload to drivers even if that speed caused anxiety to driver. Drivers who have somewhat driving experience are familiar with driving situation and speed. In the real way, especially, which in the city in Korea, the average driving speed is 40km/h and accelerated average driving speed is 60km/h. The experiment conditions set 40km/h and 60km/h because of road situation, but these conditions are very familiar to experienced drivers. So, experienced drivers may didn't need to use quite amount of mental workload which allocated in Central executive resources. And second, because of that familiarity, anxiety also not caused which could affect to mental workload enough even the experienced drivers were reported that they felt anxiety.

But when we care about driving experience, it accorded with expected result. Except one participant, novice drivers' pupil distance was shortened when the driving speed went higher which seems like visual tunneling phenomenon was happened. In this part, we can explain this phenomenon according to previous studies and our interpretation. Novice drivers are not familiar with driving situation, so they need mental workload more to understand their situation and cope with the changeable environment. Even more, novice driver felt anxious feeling when they driving in higher speed, so anxiety caused more mental workload and then it affected to allocation of Central executive resources because of much loaded mental workload. From this process, novice driver's visual attention was influenced by speed.

And there was considerable thing was observed. When the drivers were focused on the near target at 60km/h, driver's pupil distance was unexpectedly widened. We considered it caused by unconscious safety protective instinct from this phenomenon. One subjects was told, when driving speed went high while driver's focused on the near target, she felt anxiety because of she could not focus on the fore seen which can give the information for keeping safety but she tried to staring target continuously. Following the natural human eye movement, pupil distance is shortened when human focus on the near object and widened when human focus on the far object. In this experiment, as speed continues to increase, along with pupil distance while subject focus on the same distance target. So, we interpreted that even if they focused on the same distance target, they want to get information of fore seen for their safety unconsciously so that there can be some mechanism.

6 Conclusion

In conclusion, we note the difference among drivers in their efficacy of processing the visual information acquired while driving due to the difference in their level of experience/expertise in driving. This difference in efficacy incurs the difference in the maximum amount of visual information individual drivers can take/handle while driving.

On the other hand, drivers tend to take a farther look as their driving speeds increase, so as to have enough time for dealing with potential hazards, if any. This implies that the increase in driving speed heightens the amount of visual information for a driver to process due not only to the increased input stream but also to the increased field of view with a farther look, thus burdening the mental workload of a driver. The mental workload of a driver could be further exacerbated as the speed related anxiety stress starts to be accumulated. Especially, the anxiety stress of a driver becomes worsened significantly, when his/her driving speed causes the amount of visual information to process to exceed the maximum amount of visual information he/she can take/handle.

Based on the above observations, we can now analyze our experimental results on how the driving speed affects the visual attention of drivers, as follows:

For the sake of ensuring safety, drivers tend naturally to take as much visual information as possible, should it be allowed in terms of the maximum amount of visual information they can take/handle. However, as mentioned above, the maximum visual information a driver can take/handle is different among individuals according to their level of driving expertise. Since the increase of driving speed increases the amount of visual information to process, a driver may be able to expand their field of view up to the point where the amount of visual information to process at the very driving speed balances the maximum amount he/she can take/handle under the increased anxiety stress taken into consideration. Should the driving speed breaks over this balance point, a driver intend to self-maintain the balance point by automatically narrowing down the field of view, thus reducing the amount of visual information to process. For certain drivers, the anxiety stress over the balance point is so high that the maximum visual information they can handle after breaking the balance point is even further diminished.

Acknowledgement. This research was supported by Basic Science Research Program through NRF of Korea, funded by MOE(NRF-2010-0020210), by the National Research Foundation of Korea(NRF) Grant funded by the Korean Government(MSIP) (NRF-2012R1A1A3008188), and by This research was supported by MSIP, Korea under ITRC NIPA-2013-(H0301-13-3001). It is also partially supported by MEGA science research and development projects, funded by Ministry of Science, ICT and Future Planning (2013M1A3A3A02042335). We appreciate Hyekeong Kim and Seongho Cho for their assistance in experimentation.

References

1. Recarte, M.A., Nunes, L.M.: Mental workload while driving: effects on visual search, discrimination, and decision making. Journal of Experimental Psychology: Applied 9(2), 119 (2003)

2. Recarte, M.A., Nunes, L.M.: Effects of verbal and spatial-imagery tasks on eye fixations while driving. Journal of Experimental Psychology: Applied 6(1), 31 (2000)
3. Recarte, M.A., Nunes, L.: Mental load and loss of control over speed in real driving.: Towards a theory of attentional speed control. Transportation Research Part F: Traffic Psychology and Behaviour 5(2), 111–122 (2002)
4. Fu, R., Guo, Y., Chen, Y., Yuan, W., Ma, Y., Peng, J., Wang, C.: Research on Heart Rate and Eye Movement as Indicators of Drivers' Mental Workload. In: 3rd International Conference on Road Safety and Simulation (2011)
5. Liu, N., Zhang, K., Sun, X.: The Measurement of Driver's Mental Workload: A Simulation-Based Study. In: International Conference on Transportation Engineering 2007, pp. 1187–1193. ASCE (2007)
6. Wong, J.T., Huang, S.H.: Modeling Driver Mental Workload for Accident Causation and Prevention. Journal of the Eastern Asia Society for Transportation Studies 8, 1918–1933 (2009)
7. Eysenck, M.W., Calvo, M.G.: Anxiety and performance: The processing efficiency theory. Cognition & Emotion 6(6), 409–434 (1992)
8. Masters, R.S.W., Polman, R.C., Hammond, N.V.: 'Reinvestment': A dimension of personality implicated in skill breakdown under pressure. Personality and Individual Differences 14(5), 655–666 (1993)
9. Lavric, A., Rippon, G., Gray, J.R.: Threat-evoked anxiety disrupts spatial working memory performance: an attentional account. Cognitive Therapy and Research 27(5), 489–504 (2003)
10. Shackman, A.J., Sarinopoulos, I., Maxwell, J.S., Pizzagalli, D.A., Lavric, A., Davidson, R.J.: Anxiety selectively disrupts visuospatial working memory. Emotion 6(1), 40 (2006)
11. Wilson, M., Smith, N.C., Chattington, M., Ford, M., Marple-Horvat, D.E.: The role of effort in moderating the anxiety–performance relationship: Testing the prediction of processing efficiency theory in simulated rally driving. Journal of Sports Sciences 24(11), 1223–1233 (2006)
12. Matthews, M.L., Bryant, D.J., Webb, R.D., Harbluk, J.L.: Model for situation awareness and driving: Application to analysis and research for intelligent transportation systems. Transportation Research Record: Journal of the Transportation Research Board 1779(1), 26–32 (2001)
13. Easterbrook, J.A.: The effect of emotion on cue utilization and the organization of behavior. Psychological Review 66(3), 183 (1959)
14. Derryberry, D., Tucker, D.M.: Motivating the focus of attention (1994)
15. Friedman, R.S., Förster, J.: Implicit affective cues and attentional tuning: an integrative review. Psychological Bulletin 136(5), 875 (2010)
16. Janelle, C.M., Singer, R.N., Williams, A.M.: External distraction and attentional narrowing: Visual search evidence. Journal of Sport & Exercise Psychology (1999)
17. Spielberger, C.D.: Manual for the State-Trait Anxiety Inventory STAI (Form Y) ("Self-Evaluation Questionnaire") (1983)
18. Dula, C.S., Adams, C.L., Miesner, M.T., Leonard, R.L.: Examining relationships between anxiety and dangerous driving. Accident Analysis & Prevention 42(6), 2050–2056 (2010)
19. Marteau, T.M., Bekker, H.: The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). British Journal of Clinical Psychology 31(3), 301–306 (1992)

Predicting Eyes' Fixations in Movie Videos: Visual Saliency Experiments on a New Eye-Tracking Database

Petros Koutras, Athanasios Katsamanis, and Petros Maragos

School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece 15773
{pkoutras,nkatsam,maragos}@cs.ntua.gr

Abstract. In this paper we describe the newly created eye tracking annotated database *Eye-Tracking Movie Database ETMD* and give some preliminary experimental results on this dataset using our new visual saliency frontend. We have developed a database with eye-tracking human annotation that comprises video clips from Hollywood movies, which are longer in duration than the existing databases' videos and include more complex semantics. Our proposed visual saliency frontend is based on both low-level features, such as intensity, color and spatio-temporal energy, and face detection results and provides a single saliency volume map. The described new eye-tracking database can become useful in many applications while our computational frontend shows to be promising as it gave good results on predicting the eye's fixation according to certain metrics.

Keywords: Eye-tracking Database, Visual Saliency, Spatio-Temporal Visual Frontend, 3D Gabor Filters, Lab Color Space.

1 Introduction

Visual attention is a mechanism employed by biological vision systems for selecting the most salient spatio-temporal regions from a visual stimuli. Attention may have two modes, a top-down task-driven, and a bottom-up data-driven, and so there is often a confusion between attention and visual saliency, which is a bottom-up process and is based on low level sensory cues of a given stimulus. On the other hand, visual attention includes many high level topics, such as semantics, memory, object searching, task demands or expectations.

The development of computational frameworks that model visual attention is critical for designing human-computer interaction systems, as they can select only the most important regions from a large amount of visual data and then perform more complex and demanding processes. Attention models can be directly used for movie summarization, by producing video skims, or by constituting a visual frontend for many other applications, such as object and action recognition. Eyes' fixation prediction over different stimulus appeared to be a

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 183–194, 2014.

widely used way for analyzing and evaluating visual attention models. Although many databases with eye tracking data are available [1], most of them contain only static images, as the first saliency models were based only on static cues.

The purpose of this paper is to describe the newly created eye tracking annotated database *Eye-Tracking Movie Database ETMD* and give some preliminary experimental results on this dataset using our new visual attention frontend, that is based on both low level streams and mid-level cues (i.e. face detection). The existing eye-tracked video databases, in most cases contain very short videos with simple semantic contain. In our effort to deal with more complex problems, such as movie summarization [2], we have developed a database with eye-tracking human annotation, which comprises video clips from Hollywood movies, which are longer in duration and include more complex semantics.

In the second part of the paper, we describe ways for predicting the eye's fixations in movie videos and give preliminary results from our computational framework for visual saliency estimation. Our proposed visual saliency frontend is based on both low-level features [3–5], such as intensity, color and motion, and face detection results and provides a single saliency volume map. We quantitatively evaluate our results according to 3 evaluation scores, as they are described in [6]: Correlation Coefficient, Normalized Scanpath Saliency, Shuffled Area Under Curve. The described new eye-tracking database can become useful in many applications while our computational frontend shows to be promising as it gave good results on predicting the eye's fixation according to all three employed metrics.

2 Eye-Tracking Movie Database (ETMD)

We have developed a new database comprising video clips from Hollywood movies which we have enriched with eye-tracking human annotation: the Eye-Tracking Movie Database (ETMD). Specifically, we cut 2 short video clips (about 3-3.5 minutes) from each one of six Oscar-winning movies of various genres: Chicago (CHI), Crash (CRA), Departed (DEP), Finding Nemo (FNE), Gladiator (GLA) , Lord of the Rings - The return of the King (LOR). We have tried to include scenes with high motion and action as well as dialogues. These clips were annotated with eye-tracking data by 10 different people (annotation data from at least 8 people were collected for each clip). The volunteers viewed the videos both in grayscale and in color, while an eye-tracking system recorded their eyes fixations on the screen.

Specifically, we have used the commercial Eye Tracking System TM3 provided by EyeTechDS. This device uses a camera with infrared light and provides a real time continuous gaze estimation, defined as fixation points on the screen. The tracker's rate has been limited by the video frame rate in order to have one fixation point pair per frame. For a visual attention problem a weighted average between two eye fixations is provided, which is defined either by the mean, if both eyes are found by the eye-tracker, or only by the detected eye's fixation. If neither eye is detected or the fixations lie out of screen boundaries, fixation gets a

(a) CHI Clip 1 (b) CHI Clip 2 (c) CRA Clip 1 (d) CRA Clip 1

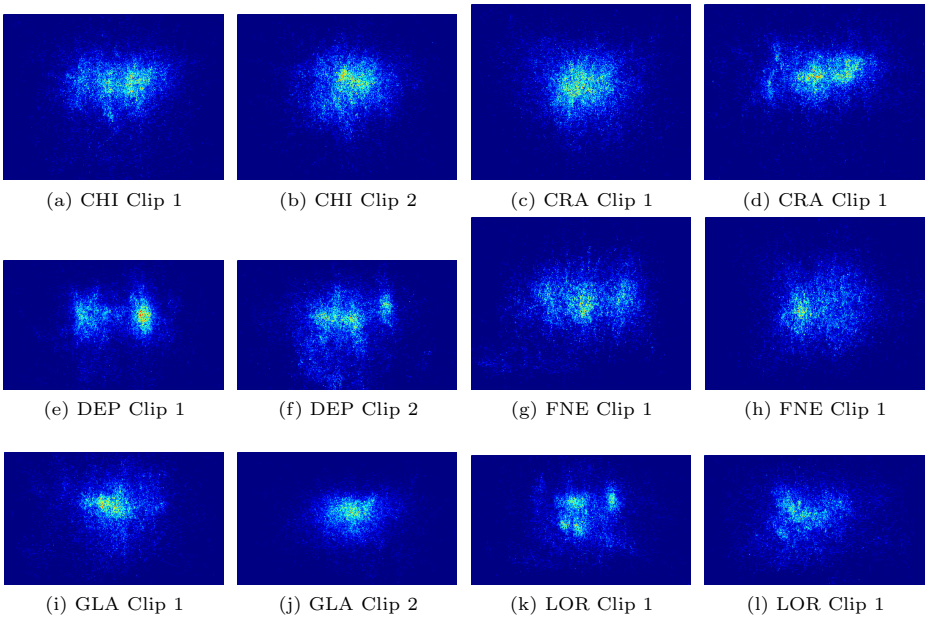
(e) DEP Clip 1 (f) DEP Clip 2 (g) FNE Clip 1 (h) FNE Clip 1

(i) GLA Clip 1 (j) GLA Clip 2 (k) LOR Clip 1 (l) LOR Clip 1

Fig. 1. Examples of the fixation points at frame no. 500 for each of the 12 movie clips. With green + are the fixations points over the color version of each clip, while with red ∗ are the points for the grayscale version. Best viewed in color.

zero value. The eye-tracking system also provides some additional measurements, such as pupil and glints positions and pupil diameter.

Figure 1 shows examples of the fixation points at frame no. 500 for each of the 12 movie clips. We see that in most cases the fixation points of all viewers lie in general close to each other. The fixations for the grayscale version of each clip are highly correlated with the fixation points over the color video as well. Figure 2 shows heatmaps of all fixations over each movie clip for the ETMD database. We see that the most points are clustered at the center of the image which shows that movie clips are highly center-biased. Analyzing the eye-tracking data we provide in Table 1 useful statistics for the database, such as frames number, total duration and valid fixation points per frame, and find correlations among the different viewers and between the color and grayscale version of each movie clip. We see that the fixations are generally correlated both between the different users and the version (color or grayscale) of each movie clip. However, in some movies, such as CHI, the fixations data are highly correlated while other clips (FNE Clip 2, LOR Clip 2) have lower correlation values.



(a) CHI Clip 1 (b) CHI Clip 2 (c) CRA Clip 1 (d) CRA Clip 1

(e) DEP Clip 1 (f) DEP Clip 2 (g) FNE Clip 1 (h) FNE Clip 1

(i) GLA Clip 1 (j) GLA Clip 2 (k) LOR Clip 1 (l) LOR Clip 1

Fig. 2. Heatmaps of all fixations over each movie clip for the ETMD database. We can see that the most points are clustered at the center of the image which shows that movie clips are highly center-biased. Best viewed in color.

Table 1. Statistics for the Eye Tracking Movie Database

Video Clip Name	Number of Frames	Duration (Minutes)	Number of Viewers	Valid Fixations Number per Frame	Average Correlation between Viewers	Average Correlation between Color and Grayscale version
CHI Clip 1	5075	03:22	10	9.50	0.506	0.495
CHI Clip 2	5241	03:29	9	8.63	0.430	0.484
CRA Clip 1	5221	03:28	10	9.47	0.335	0.310
CRA Clip 2	5079	03:23	9	8.47	0.406	0.467
DEP Clip 1	4828	03:13	10	9.45	0.520	0.548
DEP Clip 2	5495	03:39	9	8.25	0.473	0.534
FNE Clip 1	5069	03:22	9	8.45	0.372	0.371
FNE Clip 2	5083	03:23	8	7.50	0.292	0.294
GLA Clip 1	5290	03:31	9	8.18	0.423	0.407
GLA Clip 2	4995	03:19	8	7.61	0.354	0.443
LOR Clip 1	5116	03:24	9	8.38	0.452	0.431
LOR Clip 2	5152	03:26	8	7.56	0.294	0.283

3 Spatio-Temporal Framework for Visual Saliency

3.1 Overall Process

Our proposed visual saliency frontend is based on both low-level features [3–5], such as intensity, color and motion, and face detection results and provides a single saliency volume map. The overall process is shown in Fig. 3. In the first phase the initial RGB video volume is transformed into Lab space [7] and split into two streams: luminance and color contrast. For the luminance channel

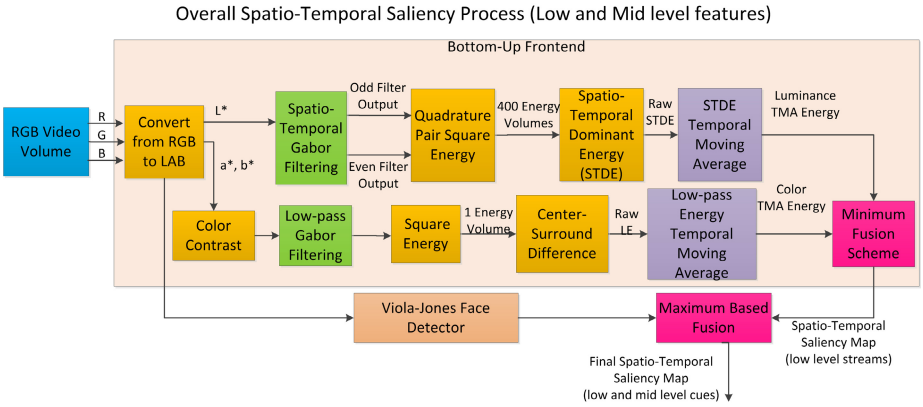


Fig. 3. Overall process for spatio-temporal saliency estimation, which includes both low-level features (i.e. intensity, color and motion) and face detection results and provides a single saliency volume map

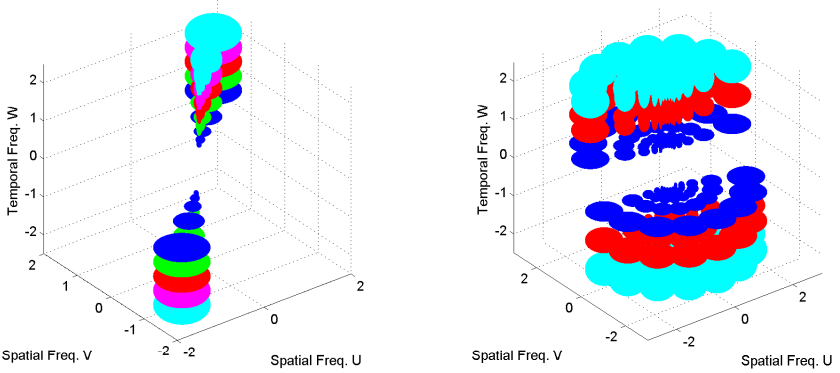
we apply spatio-temporal Gabor filtering [8, 9], followed by Dominant Energy Selection [10, 11], while for the color contrast stream we apply a simple lowpass 3D Gaussian filter followed by a center-surround difference. For integrating the results from the Viola-Jones face detector [12] to a final visual attention map we can use either a maximum or use the face detector's estimation only for the frames that contain faces.

3.2 Spatio-Temporal Bottom-Up Frontend

Preprocessing and Color Modeling. For the color modeling we use the CIE-Lab color space because in this space luminance and chromaticity components can be well separated while it has the additional property to be perceptually uniform. The CIE-Lab space is created from a nonlinear transformation on CIE-XYZ color space [13]. Then the three CIE-Lab components (L^*, a^*, b^*) can be computed by a non-linear transformation of the CIE tristimulus values (X, Y, Z). In the resulting video volume $\mathbf{I}_{Lab}(x, y, t)$ the L^* component expresses the perceptual response to luminance, while a^*, b^* describe differences between red-green and yellow-blue colors respectively. So, the CIE-Lab space includes ideas from the color-opponent theory, which was widely used in visual saliency models and was usually implemented in the RGB color space [5]. In order to describe the color changes in videos by a single measure with positive values, we use the following color contrast operator based on the chromaticity components (a^*, b^*):

$$C_{ab}(x, y, t) = |a^*(x, y, t)| + |b^*(x, y, t)| \tag{1}$$

3D Gabor Filtering. For the filtering process of the video's luminance we choose to use oriented Gabor filters in a spatio-temporal version, due to their



(a) Spatio-Temporal Filterbank at 5 different spatial scales, 1 of 8 orientation and 5 temporal frequencies

(b) Spatio-Temporal Filterbank at 5 different spatial scales, 8 spatial orientations and 3 of 5 temporal frequencies

Fig. 4. Isosurfaces of the 3D Spatio-Temporal Filterbank. Isosurfaces correspond at 70%-peak bandwidth magnitude while different colors are used for different temporal frequencies. We can see that the symmetric lobe of each filter appeared at the plane defined by the temporal frequency $-\omega_{t_0}$ in contrast with the 2D case. We also note that the bandwidth of each filter changes depending on the spatial scale and temporal frequency.

biological plausibility and their uncertainty-based optimality [14,15]. Specifically, we apply quadrature pairs of 3D (spatio-temporal) Gabor filters with identical central frequencies and bandwidth. These filters can arise from 1D Gabor filters [14] in a similar way as Daugman proposed 2D Oriented Gabor Filters [16]. An 1D complex Gabor filter consists of a complex sine wave modulated by a Gaussian window. Its impulse response with unity norm has the form:

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) \exp(j\omega_{t_0} t) = g_c(t) + j g_s(t) \tag{2}$$

The above complex filter can be split into one odd(sin)-phase ($g_s(t)$) and one even(cos)-phase ($g_c(t)$) filters, which form a quadrature pair filter.

The 3D Gabor extension (as for example used for optical flow in [9]) yields an *even (cos)* 3D Gabor filter whose impulse response is:

$$g_c(x,y,t) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_t} \exp\left[-\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2} + \frac{t^2}{2\sigma_t^2}\right)\right] \cdot \cos(\omega_{x_0} x + \omega_{y_0} y + \omega_{t_0} t) \tag{3}$$

where $\omega_{x_0}, \omega_{y_0}, \omega_{t_0}$ are the spatial and temporal angular center frequencies and $\sigma_x, \sigma_y, \sigma_t$ are the standard deviations of the 3D Gaussian envelope. Similarly for the impulse response of *odd (sin)* filter which we denote by $g_s(x,y,t)$.

The frequency response of the even (cos) 3D Gabor Filter will have the form:

$$\begin{aligned} G_c(\omega_x, \omega_y, \omega_t) = & \frac{1}{2} \exp[-(\sigma_x^2(\omega_x - \omega_{x_0})^2/2 \\ & + \sigma_y^2(\omega_y - \omega_{y_0})^2/2 + \sigma_t^2(\omega_t - \omega_{t_0})^2/2)] \\ & + \frac{1}{2} \exp[-(\sigma_x^2(\omega_x + \omega_{x_0})^2/2 \\ & + \sigma_y^2(\omega_y + \omega_{y_0})^2/2 + \sigma_t^2(\omega_t + \omega_{t_0})^2/2)] \end{aligned} \quad (4)$$

Thus, the frequency response of an even (cos) Gabor filter consists of two Gaussian ellipsoids symmetrically placed at frequencies $(\omega_{x_0}, \omega_{y_0}, \omega_{t_0})$ and $(-\omega_{x_0}, -\omega_{y_0}, -\omega_{t_0})$. Figure 4 shows isosurfaces of the 3D spatio-temporal filterbank. Note that the symmetric lobes of each filter appear at the plane defined by the temporal frequency $-\omega_{t_0}$ in contrast with the 2D case. So, if we want to cover the spatial frequency plane at each temporal frequency we must include in our filterbank both positive and negative temporal frequencies. Further, the bandwidth of each filter varies with the spatial scale and temporal frequency.

The 3D filtering is a time consuming process due to the complexity of all required 3D convolutions. However, Gabor filters are separable [9], which means that we can filter each dimension separately using an impulse response having the form (2). In this way, we apply only 1D convolutions instead of 3D, which increases the efficiency of the computations. For an image of size $n \times n \times n$ and a convolution kernel of $m \times m \times m$ the complexity is reduced from $\mathcal{O}(n^3 \cdot m^3)$ that is required for 3D convolutions to $\mathcal{O}(3n^3 \cdot m)$ that is required for three separable 1D convolutions.

For the spatio-temporal filterbank we used $K = 400$ Gabor filters (isotropic in the spatial components) which are arranged in five spatial scales, eight spatial orientations and ten temporal frequencies. The spatial scales and orientations are selected to cover a squared 2D frequency plane in a similar way to the design by Havlicek et al. [11]. We also use ten temporal Gabor filters, five at positive and five at negative center frequencies due to the 3D spectrum symmetries. Figure 4 shows spatio-temporal views of our design of this 3D filterbank.

Finally, for the low-pass color filtering we use both spatial and temporal zero frequencies which makes the Gabor filter gaussian.

Postprocessing. After the filtering process, for each filter i we obtain a quadrature pair output $(y_s^{3D}(x, y, t), y_c^{3D}(x, y, t))$ which corresponds to the even- and odd-phase 3D filter outputs. We can compute the total Gabor energy, which is invariant to the phase of the input, by taking the sum of the squared energy of these two outputs:

$$STE_i(x, y, t) = \left(y_s^{3D}(x, y, t)\right)^2 + \left(y_c^{3D}(x, y, t)\right)^2 \quad (5)$$

After this step we have 400 *energy volumes* for the spatio-temporal part (STE_i) and one for the lowpass color filter (LE_0). In order to form one volume for the luminance modality we apply the first step of *Dominant Component Analysis* to

spatio-temporal volumes. Specifically, for each voxel (x, y, t) we keep its maximum value between all existing energy volumes: $STDE = \max_{1 \leq i \leq K} STE_i$ For the lowpass color energy we apply a simple center-surround difference in order to enhance regions which have significantly different values from their background. At each voxel of the video segment we subtract from its lowpass energy value $(LE_0(x, y, t))$ the mean value of the entire energy volume:

$$LE(x, y, t) = \left| LE_0(x, y, t) - \overline{LE}_0(x, y, t) \right| \tag{6}$$

Finally, these energy volumes can become further smoothed by applying a *temporal moving average* (TMA). Thus, each frame energy is computed as the mean inside a temporal window which includes N successive frames whose total duration is 1 second. In this way, we integrate visual events which take place close in time, in a similar way that humans are believed to do. A spatial smoothing with a dilation operator can also by applied, in order to find more compact and dense energy regions.

4 Evaluation on ETMD

4.1 Evaluation Measures

We have tried to keep the same evaluation framework as in [6]. We compared our results according to the three evaluation scores, as they are described in [6]: Correlation Coefficient, Normalized Scanpath Saliency, Area Under Curve. Despite the spatio-temporal character of our method these three measures are computed at each frame separately.

Correlation Coefficient (CC) expresses the relationship between the model's saliency map and the saliency map created by centering a 2D gaussian, with standard deviation 10 pixels, at each viewer's eye fixation.

Normalized Scanpath Saliency (NSS) is computed on the model's saliency map, after zero mean normalization and unit standardization, and shows how many times over the whole map's average is the model's saliency value at each human fixation. For NSS computation we subtract from the saliency map its average value and then divide with its standard deviation. Then the values of this normalized saliency map at each viewer fixation position consist the NSS values. As final NSS value we take the mean over all viewers fixations, while a negative NSS shows that the model cannot predict saliency region better than random selection.

Area Under Curve (AUC) is defined by the area under the Receiver Operating Characteristic (ROC) curve [17]. For our evaluation we consider saliency as a binary classification problem, in which saliency regions are included in the positive class while non-salient pixels form the negative set and model's saliency values are the single features. After thresholding these values we take an ROC curve and subsequently the AUC measure. Instead of selecting the negative points uniformly from a video frame we use the *Shuffled AUC*, which can be more robust across center-bias issue. According to shuffled AUC, we select the negative points

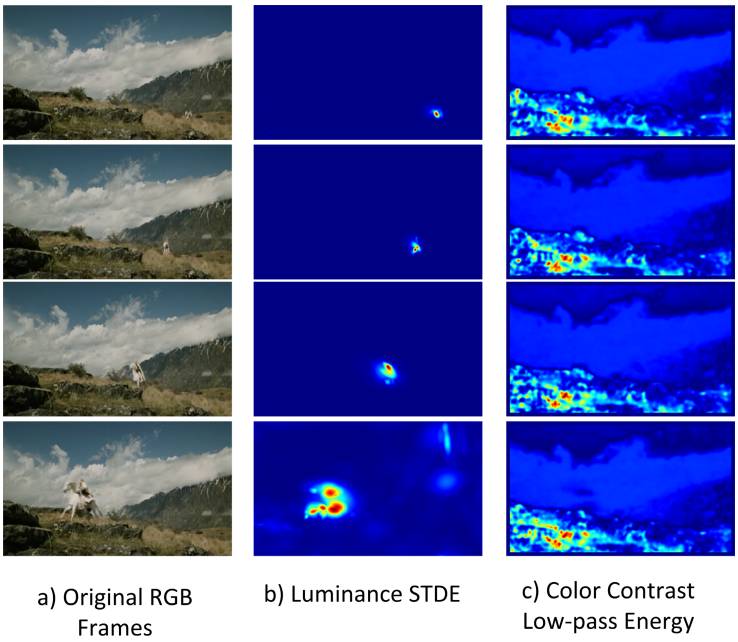


Fig. 5. Example frames of energy volumes computed using our frontend on the *Lord of the Rings* (Clip 1) from our Eye-Tracking Movie Database (ETMD). The galloping horse is perfectly detected by the luminance STDE.

from the union of all viewers' fixations across all other videos except the video for which we compute the AUC. For more details about the above evaluation scores the reader is referred to [1, 6].

4.2 Evaluation Results

We have applied and evaluated our computational model on this novel database. Figure 5 shows example frames of the model's energies computed on the video *Lord of the Rings (LOR)* (Clip 1) from our new Eye-Tracking Movie Database (ETMD). We note that the white galloping horse is detected perfectly by only the luminance $STDE$, since its color information is negligible. The color low-pass energy models static objects or regions in the video sequence, like the rock in the bottom-left and the clouds in the air.

Regarding the feature energy volumes employed we see that luminance $STDE$ and the color low-pass energy using a min fusion performs quite better than using only the $STDE$ energy volume. Moreover, Finally, regarding the grayscale versus color annotation, we saw that the evaluation over color videos yields better results, which indicates that the way color attracts human attention may be predicted more accurately by our model.

Table 2. Evaluation Scores for the Eye-Tracking Movie Database(ETMD) using our **bottom-up frontend**. The employed evaluation measures are Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) and Shuffled Area Under Curve (AUC). The evaluation of the Luminance $STDE$ was based on Eye-Tracking annotation on both a grayscale and color version of each video.

Evaluation Score	Correlation Coefficient (CC)	Normalized Scanpath Saliency (NSS)	Shuffled Area Under Curve (AUC)
Lum. $STDE$ (Grayscale Annot.)	0.151	0.608	0.611
Lum. $STDE$ (Color Annot.)	0.153	0.632	0.614
MIN(Lum. $STDE$, Color Low-pass)	0.169	0.748	0.635

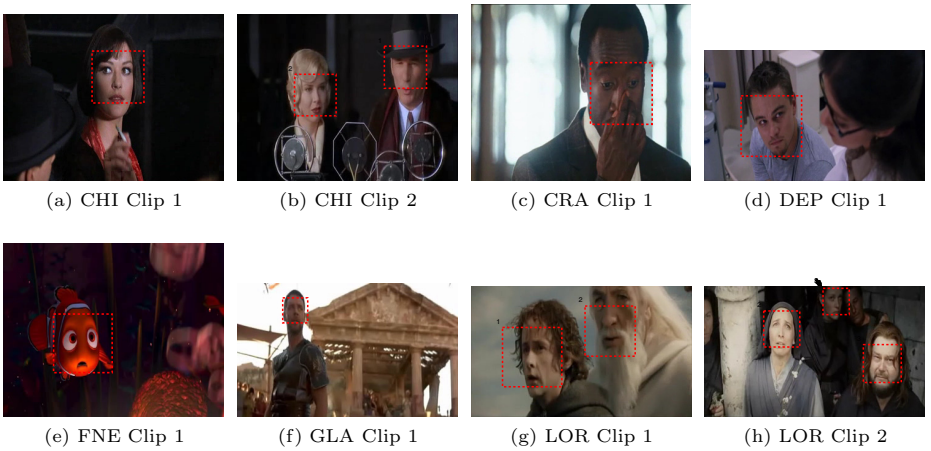
From the database analysis we have seen that in many cases the humans have focused on actors' faces, while their eyes' fixation has also the trend to be center biased. To model these two effect we use two simple methods. The first consists of using a gaussian kernel fixed at the image kernel. The latter provides the use of the Viola-Jones face detector as a saliency estimator only in the frames where people face exist. We have also tried to predict one viewers fixations from the other users eye fixations, as reference results.

4.3 Face Detection

Figure 6 shows examples of the Viola-Jones face detector [12] applied on videos from the new ETMD. We see that in many cases the human faces have been detected accurately while in Fig. 6e the detector find the human-like face of the fish "Nemo". We note that the employed face detector is not very robust with changes in face pose and scale. Thus, it cannot achieve to detect all the faces during a video (low recall) but the obtained results are in most cases true (high precision).

For the fusion of the visual saliency estimation from our bottom-up frontend with the face detection results we can use either a max based method (MAX(Bottom-Up, Face Detection)) or by using the face detector's result for the frames with faces and the BU model for the other frames (Bottom-Up OR Face Detection). We have also applied the face detector and bottom-up model independently only for the frames that contain faces. Table 3 shows the results related with the use of the face detector. We see that face detector improves the final visual saliency result, especially in frames that contain faces, while the two fusion methods' results are very closely.

In this Table are also presented the scores achieved by a Gaussian blob centered at the center of the image as well as the results related with the prediction of each one viewer's fixations from the other users eye fixations. We can see that the Gaussian blob gave just as good results w.r.t. CC and NSS, which confirms the existence of the center-bias effect. Regarding the shuffled AUC it has lower performance since this measure is more suitable for high center-biased database. Regarding the fixation prediction from the other viewers' data, it has achieved high performance w.r.t. CC and NSS measures because even in few occasions humans look at the same direction, this hits very large NN and CC values.



(a) CHI Clip 1 (b) CHI Clip 2 (c) CRA Clip 1 (d) DEP Clip 1

(e) FNE Clip 1 (f) GLA Clip 1 (g) LOR Clip 1 (h) LOR Clip 2

Fig. 6. Examples of Face Detection results over the movie clips from the ETMD database. Best viewed in color.

Table 3. Evaluation Scores for the Eye-Tracking Movie Database(ETMD) using our **bottom-up frontend and the Viola-Jones face detector**. The employed evaluation measures are Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) and Shuffled Area Under Curve (AUC). There are also presented the scores achieved by the Gaussian blob and the prediction of each one viewer's fixations from the other users eye fixations.

Evaluation Score	Correlation Coefficient (CC)	Normalized Scanpath Saliency (NSS)	Shuffled Area Under Curve (AUC)
Face Detection Only	0.327	1.622	0.807
Bottom-Up (only frames with faces)	0.164	0.719	0.636
Bottom-Up OR Face Detection	0.203	0.933	0.680
MAX(Bottom-Up, Face Detection)	0.201	0.919	0.680
Gaussian Blob	0.197	1.288	0.580
Predict from other viewers' fixations	0.404	2.515	0.617

5 Conclusion

In this paper we presented a new eye-tracking database which comprises video clips from Hollywood movies and eye-tracking data recorded from different viewers. We have also given evaluation results using our proposed spatio-temporal bottom-up frontend for visual saliency estimation. We have also dealt with the problem of "face-biased" movie video by combining the results from our bottom-up saliency frontend with a face detector's estimation. We believe that both the new eye-tracking database and our framework for predicting eye fixations can become useful in many computer applications, such as the producing of movie summaries.

Acknowledgment. The authors wish to thank all the members of the NTUA CVSP Lab who participated in the eye-tracking annotation of the movie database.

This research was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund (ESF) and National Resources.

References

1. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Analysis and Machine Intelligence 35(1), 185–207 (2013)
2. Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., Avrithis, Y.: Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention. IEEE Trans. on Multimedia 15(7) (2013)
3. Treisman, A., Gelade, G.: A feature integration theory of attention. Cognit. Psychology 12(1), 97–136 (1980)
4. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4(4), 219–227 (1985)
5. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
6. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Trans. Image Processing 22(1), 55–69 (2013)
7. Poynton, C.: Digital Video and HD: Algorithms and Interfaces, 2nd edn. Morgan Kaufmann (2012)
8. Adelson, E.H., Bergen, J.R.: Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Amer. A 2(2), 284–299 (1985)
9. Heeger, D.J.: Model for the extraction of image flow. J. Opt. Soc. Amer. 4(8), 1455–1471 (1987)
10. Bovik, A.C., Gopal, N., Emmoth, T., Restrepo, A.: Localized Measurement of Emergent Image Frequencies by Gabor Wavelets. IEEE Trans. Information Theory 38, 691–712 (1992)
11. Havlicek, J.P., Harding, D.S., Bovik, A.C.: Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models. IEEE Trans. Image Processing 9(2), 227–242 (2000)
12. Viola, P., Jones, M.J.: Robust real-time face detection. Int'l. J. Comput. Vis. 57(2), 137–154 (2004)
13. Wyszecki, G., Stiles, W.S.: Color Science, 2nd edn. J. Wiley & Sons, NY (1982)
14. Gabor, D.: Theory of Communication. IEE Journal (London) 93, 429–457 (1946)
15. Daugman, J.: Uncertainty Relation for Resolution in Space, Spatial Frequency and Orientation Optimized by Two-Dimensional Visual Cortical Filters. J. Opt. Soc. Amer. A 2(7), 1160–1169 (1985)
16. Daugman, J.G.: Two-dimensional spectral analysis of cortical receptive field profiles. Vision Research 20(10), 847–856 (1980)
17. Green, D.M., Swets, J.A.: Signal detection theory and psychophysics. Wiley, New York (1966)

The Time Course of Selective Consolidation on Visual Working Memory

Haifeng Li[1,2], Yanan Chen [1,2], and Kan Zhang [1,*]

[1] Key Laboratory of Behavioral Science
Institute of Psychology, Chinese Academy of Sciences, P.R. China
[2] University of Chinese Academy of Sciences, P.R. China
{lihf,chenyn,zhangk}@psych.ac.cn

Abstract. The aim of this study is to explore the time course of the selective consolidation process. While maintaining one color, which was selected from an initially array of four colors in three different intervals, participants had to perform a visual search and ignore 1) a singleton distractor that matched the color maintained in WM, 2) a new singleton that did not match the maintained color, or 3) a uniformly colored distractor display (i.e., no singletons). After that, they should match a probe color with the maintained color. WM performance for the color was significantly impaired in the matching and new color distractor conditions relative to the uniformly distractor condition when the interval for selective consolidation was short (35 or 50 ms/item), but was identical across these three conditions when this interval was relative long (65 ms/item). This result indicated an astonishingly fast process of selection before the color consolidation process.

Keywords: time course, selective consolidation, visual working memory.

1 Introduction

Visual working memory (WM) serves to temporarily maintain and manipulate information important for behavior [1]. Besides the limited storage capacity of WM system, the limited processing capacity of consolidation also affects the WM maintenance. Consolidation is a process of encoding information into a durable WM representation that can survive further sensory interference [2-4].

In the real world, people should selectively concentrate on one aspect of the environment while ignore other information, then they may memorize them for further processing. For example, in safety inspection, screeners should detect the threatening objects from a pile of luggage, then report or record them. This can be regarded as a selective consolidation process.

How long does it take to select and form a durable representation in WM? Previous studies found that the sole consolidation process can be very slow, with estimates of approximately 500 ms in attentional blink [2] and dual-task [3] experiments. In the attentional blink paradigm, participants should detect and identify two targets in a rapid serial visual presentation sequence. Participants are difficult to report the second

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 195–201, 2014.

target (T2) when T2 is presented 200-600 ms after the first target (T1). The failure to report T2 has been regarded as a failure to consolidate T2 into a durable WM representation while T1 is going through consolidation [4, 5]. Jolicoeur and Dell'Acqua used a dual-task procedure in which combined a verbal WM task with an auditory speeded-response task [3]. Participants should remember a visual array of characters (T1), then immediately response to a successive low- or high-pitched tone (T2). At the end of the trial, memory for the character array was tested. They also found substantial slow in response to the tone when the delay between T1 and T2 was short (350-550 ms). They proposed that the performance of the auditory task was interfered when the memory array had not yet been consolidated.

But later, Vogel, Woodman, and Luck (2006) revealed that the rate of consolidation can be much shorter, approximately 50 ms per item [6]. In this study, participants performed a change-detection task for color squares. Pattern masks were presented in the locations of the squares shortly after the presentation of the memory array to disrupt the color consolidation. They found the memory performance was impaired due to the short delay between the squares and the masks. Does the selective consolidation have the similar time course as the sole consolidation process? Or how long will the selection process take before consolidation?

In our one previous study [7], participants maintained one color selected from an array of two or four colors, and then passively viewed a distractor display among which they had to ignore:

1. a singleton distractor that matched the color maintained in WM
2. a new singleton that did not match the maintained color, or
3. a uniformly colored distractor display (no singletons).

At the end of a trial, they should match a color with their maintained color. In this study, the duration of cue which instructed participants to remember only one color from two or four colors was 150 ms. The results found that, WM performance for single items encoded from two colors was identical across the three visual distractor display conditions. However, WM performance for single items encoded from four colors was impaired by the presence of a matching color distractor, improving further when the color distractor was new, and best when there was no singleton distractor. In another experiment, participants still selected to remember one color from four colors, but we inserted a 1000 ms blank interval between the offset of the cue and the onset of the distractor display. The result revealed that WM performance was no differences between the three visual distractor display conditions. Thus, if we assume that consolidation is processed in serial [6, 8], a color representation can be selectively consolidated within 75 ms (150 ms / 2) or 250 ms (1000 ms / 4), but not within 38 ms (150 ms / 4).

In the present study, to precisely measure the time course of selective consolidation, we varied the duration of cue which instructed subjects to remember only one color from four colors: 140 ms, 200 ms, or 260 ms (35 ms/item, 50 ms/item, or 65 ms/item respectively). If the selective consolidation process was complete, then singleton distractor in the distractor display will not affect the color maintenance; if not, WM performance of a single color will differ among different distractor types in the distractor display.

2 Method

2.1 Participants

Twenty-two undergraduate and graduate students were paid for their participation. Two participants were excluded due to low memory accuracy (less than 80%). Thus a total of 20 participants (17 females; 19-30 years) were included in the analyses. All participants had normal or corrected-to-normal vision and none reported color blindness.

2.2 Materials

The color stimuli could appear in (RGB: 250, 20, 0), green (RGB: 0, 170, 0), yellow (RGB: 220, 200, 20), blue (RGB: 0, 90, 200), pink (RGB: 210, 0, 110) or Gray (RGB: 85, 85, 85). The font was Arial. All stimuli were presented on a black background.

2.3 Apparatus

The experiment was programmed using E-prime 2.0, and was run on a 17-inch CRT monitor at a viewing distance of approximately 60 cm without a chin rest. The monitor was set to a 1280 × 1024 resolution with a 85 Hz refresh rate and 32-bit colors.

2.4 Procedure

The procedure was illustrated in Figure 1. At the beginning of each trial, two pseudo-random digits ranging from 1-9 were presented on the screen's center for 1,000 ms. Participants were required to repeat these two digits aloud throughout the trial to encourage them to rely on visual memory rather than phonological recoding to remember the color [9]. Participants' voices were recorded to make sure that they followed this articulation suppression task throughout the experiment.

Next, a 500 ms "remember" instruction appeared in gray in the center of the screen, followed by two or four different color disks (each with a radius 0.9° visual angle) that appeared for 1,000 ms, each centered in the four corners of an imaginary square (length 5° visual angle), which was displayed in the screen's center. Participants were required to remember the colors of these disks. After a 50 ms blank screen, an arrow cue (length 1.4°, height 0.8°) appeared to instruct them to maintain only the cued color.

The distractor display appeared after the arrow cue disappeared. This display consisted of a diamond (1.2° in size, presented in gray) and 7 disk distractors (each with a radius 0.6° visual angle). They were placed on the rim of an imaginary circle (radius 8°), which was centered on the fixation. An "N" or "M" (0.38° in size, presented in black) appeared within the diamond. The disk distractors contained a symbol resembling an hourglass on its side, which matched the line segments of the "N" and "M". This display will present for 1,500 ms. Participants were asked to ignore this display and wait for a probe color.

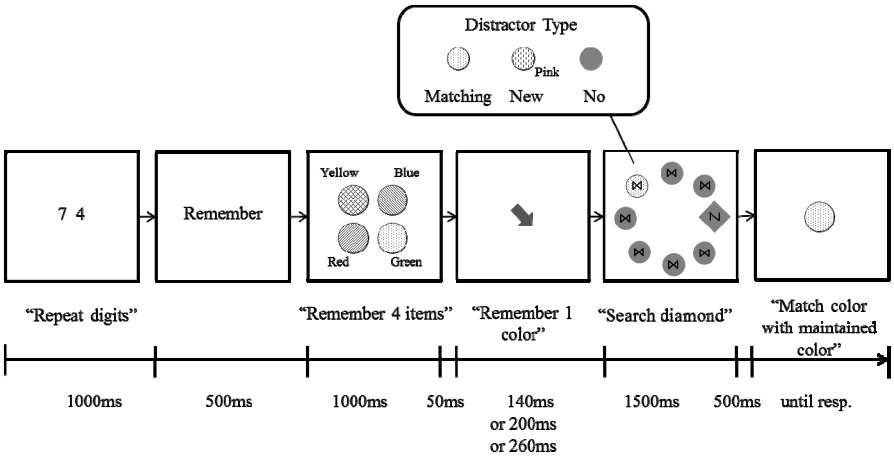


Fig. 1. Experimental procedure and example stimuli for this study. The different patterns correspond to different colors.

After another 500 ms blank screen, participants were instructed to match the color of a probe disk (radius 0.9° visual angle) that appeared in the screen's center. If the probe color matched the color they maintained, participants responded with the left arrow key; otherwise, press the right arrow key. The color matched on 1/2 of the trials, was a new probe color on 1/4 of the trials, or was an uncued color from the memory phase on 1/4 of the trials.

The duration of the cue was varied in three levels: 140 ms, 200 ms, or 260 ms (35 ms/item, 50 ms/item, or 65 ms/item respectively). Another important factor was the distractor display type, for which there were three levels. In the no color distractor condition, all of the disks were gray. In the matching color distractor condition, one of the disks appeared in the same color as the color maintained in WM (cued from the WM display). In the new color distractor condition, one of the disks appeared in a new color that was not presented during the memory phase. The diamond and color distractor in the display never appeared in adjacent positions.

2.5 Design

This was a 3 (cue duration: 140 ms, 200 ms, or 260 ms) × 3 (distractor type: matching color distractor, new color distractor, or no distractor) design experiment. There were a total of 360 trials with 40 trials per treatment combination. All the trials were randomly intermixed across the whole experiment. Each participant first received 12 practice trials, then they completed 6 blocks with 60 trials each, with a 1-min break between blocks. The whole experiment will last 40-50 minutes.

3 Results

The result was illustrated in Figure 2. A repeated measures ANOVA with cue duration type (140 ms, 200 ms, or 260 ms) and distractor type (matching color distractor, new color distractor, or no color distractor) revealed a main effect of cue duration type, $F(2, 38) = 16.93$, $MSE = .029$, $p < .001$. Pairwise comparisons using Bonferroni adjustment revealed that WM accuracy in 140 ms cue duration condition was significantly lower than in 200 ms cue duration condition ($p < .01$) and 260 ms cue duration condition ($p = .001$), also WM accuracy in 200 ms cue duration condition was lower than in 260 ms cue duration condition ($p < .01$).

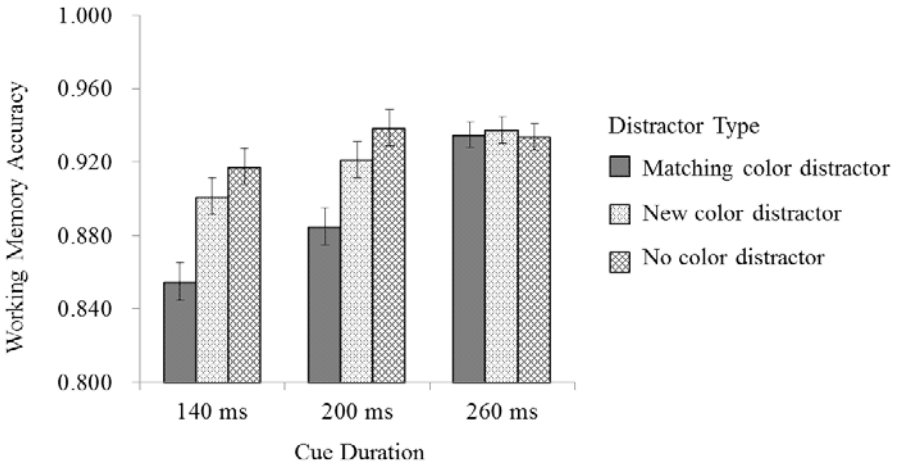


Fig. 2. WM accuracy as a function of cue duration type and distractor type. Error bars represent 95% within-participants confidence intervals with Masson and Loftus's method (2003) [10].

Also the main effect of distractor type was significant, $F(2, 38) = 7.24$, $MSE = .024$, $p < .01$. Pairwise comparisons using Bonferroni adjustment revealed that WM accuracy of the matching color distractor condition was lower than accuracy of the no color distractor conditions ($p < .01$), but there were no differences between the matching color distractor condition and the new color distractor condition ($p = .129$), and between the new color distractor condition and the no color distractor condition ($p = .693$).

Importantly, the interaction between cue duration type and distractor type was significant, $F(4, 76) = 5.90$, $MSE = .006$, $p < .001$. Simple effect analysis using Bonferroni adjustment indicated that when the cue duration was 140 ms or 200 ms, WM accuracy of the matching color distractor condition was significantly (or marginal significantly) lower than both the new color distractor condition ($p < .05$ and $p = .09$ respectively) and the no color distractor condition ($p < .01$ and $p < .001$ respectively). But there were no WM accuracy differences between the new color distractor condition and the no color distractor condition in 140 ms and 200 ms cue duration

conditions (p = .624 and p = .604 respectively). But when the cue duration was 260 ms, there were no WM accuracy differences between three distractor types (ps = 1).

4 Discussion and Conclusion

The present study showed that, when the time for selective consolidation for a single color was 35 ms/item or 50 ms/item, participants' WM performance was easily disrupted by the successive distractor display which was only exposed on the screen. But when this selective consolidation time was increased to 65 ms/item, WM performances were identical between three distractor types and showed ceiling effects.

In this study, the selection process was similar to the process that attention directed to a particular stimulus—directing attention [11]. The time course of directing attention was initially tested by Müller and Rabbitt (1989) [12]. They found a peripheral cue (presented for 50 ms), but not a central cue, could have a fast-acting effect on performance when the SOA between the cue and the search display was 100 ms. Other studies reported the similar findings [13, 14]. Cheal and Lyon (1991) systematically investigated the central and peripheral cue in a forced-choice discrimination task. They found that a peripheral cue could produce its effects on discrimination performance within 17 ms, although this effect was very weak [15]. When it comes to the time course of consolidation, in Vogel et al.'s study [6], an estimate 50 ms/item of consolidation time was derived from a single-task procedure. In the present study, we found an estimate 65 ms/item of selective consolidation time in a dual-task procedure, indicating an even faster selection process before the consolidation process.

Logically, the selection process and the consolidation process perform in serial. Participants should first chose the correct color, and then encode that color into WM. Although our result indicated that WM performance impaired when the time for selective consolidation was very short, we don't know whether the disruption occurs in the selection process or consolidation process. The similar question is, we don't know how people deploy time onto these two processes. If assuming that 50 ms/item is almost the fastest processing of consolidation, in this study, the time course for selection or directing attention will be shorter than the time previous studies have observed.

It's reasonable that WM performance is interfered by the distractors during retention. As the load theory of attention proposed that the ability to inhibit irrelevant distractors depends on higher cognitive functions, such as WM [16]. In our study, selective encoding a single color from an array of four colors was a relative high load condition, which would drain the attentional resources for attentional control and result in increased distractor interference. But a remaining question is, when the time course for selective consolidation is very short, why the memory-matching distractor in the distractor display disrupts WM more than a new color distractor or when color distractor is absent. We proposed and proved that this may be caused by the perceptual similarity or confusability between the to-be-remember item and the distractor in the distractor display [7].

This study deepens our understanding on the process of selecting and/then encoding relevant information into a temporarily WM system. It is a rather fast but fragile process. Also this finding will have implications that what one memorizes could be distorted or disrupted by what one visualizes in a very short interval, and this effect may be amplified in crime scenes, which may complicate the interpretation of eyewitness testimony [17].

In conclusion, in a dual-task procedure, the present study has shown that the rate of selective consolidation is approximately 65 ms/item. WM content is more fragile when the time for consolidation is short. The selection process before the consolidation may be considerably faster than previous suggests.

References

1. Baddeley, A.D., Hitch, G.: Working memory. In: Bower, G.A. (ed.) Recent Advances in Learning and Motivation, pp. 47–90. Academic Press, New York (1974)
2. Chun, M.M., Potter, M.C.: A two-stage model for multiple target detection in rapid serial visual presentation. J. Exp. Psychol. Hum. Percept. Perform. 21, 109–127 (1995)
3. Jolicoeur, P., Dell' Acqua, R.: The demonstration of short-term consolidation. Cognit. Psychol. 36, 138–202 (1998)
4. Shapiro, K.L., Arnell, K.M., Raymond, J.E.: The attentional blink: A review on attention and a glimpse on consciousness. Trends Cogn. Sci. 1, 291–296 (1997)
5. Vogel, E.K., Luck, S.J.: Delayed working memory consolidation during the attentional blink. Psychon. Bull. Rev. 9, 739–743 (2002)
6. Vogel, E.K., Woodman, G.F., Luck, S.J.: The time course of consolidation in visual working memory. J. Exp. Psychol. Hum. Percept. Perform. 32, 1436–1451 (2006)
7. Li, H., Chen, Y., Zhang, K.: Initial load and short consolidation time increases attentional capture and disrupts working memory (unpublished manuscript)
8. Standing, L., Haber, R.N.: Visual search and memory under degraded and masked presentation. Psychon. Sci. 13, 81–82 (1968)
9. Soto, D., Humpherys, G.W.: Stressing the mind: The effect of cognitive load and articulatory suppression on attentional guidance from working memory. Percept. Psychophys. 70, 924–934 (2008)
10. Masson, M.E.J., Loftus, G.R.: Using confidence intervals for graphically based data interpretation. Can. J. Exp. Psychol. 57, 203–220 (2003)
11. Egeth, H.E., Yantis, S.: Visual attention: Control, representation, and time course. Annu. Rev. Psychol. 48, 269–297 (1997)
12. Müller, H.J., Rabbitt, P.M.A.: Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption. J. Exp. Psychol. Hum. Percept. Perform. 15, 315–330 (1989)
13. Kröse, J., Julesz, B.: The control and speed of shifts of attention. Vision Res. 23, 1607–1619 (1989)
14. Nakayama, K., Mackeben, M.: Sustained and transient components of focal visual attention. Vision Res. 29, 1631–1647 (1989)
15. Cheal, M.L., Lyon, D.R.: Central and peripheral precuing of forced-choice discrimination. Q. J. Exp. Psychol. A. 43, 859–880 (1991)
16. Lavie, N., Hirst, A., De Fockert, J.W., Viding, E.: Load theory of selective attention and cognitive control. J. Exp. Psychol. Gen. 133, 339–354 (2004)
17. Loftus, E.F.: Eyewitness testimony. Harvard University Press, Cambridge (1979)

The Influence of Visualization on Control Performance in a Flight Simulator

Menja Scheer[1], Frank M. Nieuwenhuizen[1], Heinrich H. Bülthoff[1,2], and Lewis L. Chuang[1,⋆]

[1] Department of Perception, Cognition and Action, Max Planck Institute for Biological Cybernetics, Tübingen
[2] Department of Cognitive and Brain Engineering, Korea University
{menja.scheer,frank.nieuwenhuizen,heinrich.buelthoff, lewis.chuang}@tuebingen.mpg.de

Abstract. Flight simulators are often assessed in terms of how well they imitate the physical reality that they endeavor to recreate. Given that vehicle simulators are primarily used for training purposes, it is equally important to consider the implications of visualization in terms of its influence on the user's control performance. In this paper, we report that a complex and realistic visual world environment can result in larger performance errors compared to a simplified, yet equivalent, visualization of the same control task. This is accompanied by an increase in subjective workload. A detailed analysis of control performance indicates that this is because the error perception is more variable in a real world environment.

1 Introduction

We rely on visual feedback to ensure stable motion and collision avoidance during self-motion. Visual feedback informs the human operator of the immediate difference between his desired goal and the consequences of his action. Thus, subsequent actions can be planned to minimize this difference. For that reason, it is important to ask how error feedback should be visualized to support good control performance in the human operator.

The real world is a rich source of visual information for supporting the control of self-motion. For example, the rate and the focus of expansion in retinal image changes (i.e., optic flow) can respectively help us discern our velocity and heading direction [1,2]. Given this, it is not surprising that virtual environments often strive to achieve high visual realism. This is especially true for flight simulators that are designed to train control performance, the success of which is subsequently vital for safety in a real vehicle. Several studies support this ambition. It has been shown in a flight simulator study that increasing the realism of ground terrain results in more accurate judgments in altitude as well as improved aiming [3]. Similarly, the altitude perception in pilots improved with higher object density in the visual environment [4].

⋆ The work in this paper was supported by the myCopter project, funded by the European Commission under the 7th Framework Program.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 202–211, 2014.

Nonetheless, this strive towards high visual fidelity may not always be necessary nor helpful. For example, it has been shown in a disturbance tracking task that a simple instrument can better support the control performance of human operator than optic flow alone [5]. Similarly, a driving simulator study demonstrated that control performance is independent of whether a realistic view of the road or just the lane itself is presented [6]. Given these findings, it stands to reason that if all information necessary to complete a task could be condensed in a simple instrument, it might be possible to achieve similar performance as in a real-world environment. In fact, one might even expect better performance from a simple visualization that exclusively presents only the information that is necessary for performing a given task. This relieves the operator from parsing the environment for task relevant information.

To investigate whether or not control performance is dependent on the realism of the visualization, the present study evaluated human participants on a closed-loop control task in a high-fidelity, fixed-base flight simulator. The structure of the task is depicted in Figure 1. The reference signal $f_t(t)$ represents the target to be followed. This reference signal was an unpredictable change in the roll angle of the simulated vehicle. This was not directly shown to the *Human Operator*. Instead, only the difference $e(t)$ between the desired roll angle $f_t(t)$ and the output of the system $\varphi(t)$ was displayed. In performing this task, the *Human Operator* has to continuously perceive his deviation from the target and to manually operate a control device to minimize this perceived error.

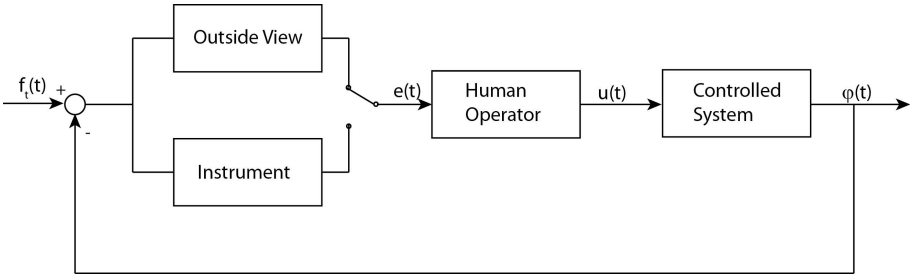


Fig. 1. Closed-loop control task of the presented study. The difference $e(t)$ between the output of the system $\varphi(t)$ and the unpredictable reference input $f_t(t)$ was presented in two different ways. Either using a simplified instrumental view (*Instrument*) or a complex visualization showing the outside view of an helicopter (*Outside View*). The human operators' task was to compensate for the disturbance introduced through the reference signal.

In our implementation of this control scheme, the *Human Operator* moved a control stick to continuously compensate for the displayed error. Moving the stick to the left or right resulted in stick deflections that were proportional to the roll rate $\dot{\varphi}(t)$ of the simulated aircraft. Thus, stick manipulations served as a direct input $u(t)$ to a *Controlled System* with single-integrator dynamics. The

output of the system was fed back and subtracted from the reference signal $f_t(t)$, resulting in the error $e(t)$ that was shown to the *Human Operator*.

As mentioned, there were the two possible visualizations for presenting this error feedback $e(t)$ to the *Human Operator*. This allowed us to investigate the influence of visualization complexity and was the only experimental manipulation in the current study. It is worth mentioning again that the reference signal $f_t(t)$ was the same regardless of the visualization. In other words, the task difficulty was the same regardless of the visualization shown. The *Instrument* visualization was comparable to an attitude indicator (commonly referred to as an artificial horizon), which is an aviation instrument that displays the aircraft's angular position with respect to the horizon (Figure 2). For the *Outside View* visualization, participants were presented with a view of a simulated real-world environment from an aircraft cockpit (Figure 3).

In the current study, we were interested in how the visualization of error feedback affected control performance. In addition, we were also motivated to know whether this influence of visualization would be accompanied by changes in subjective workload. To measure control performance, the output of the joystick $u(t)$ as well as the error $e(t)$ were recorded. $e(t)$ was the amount of error that remained in the system after the *Human Operator* resolved the continuous disturbance $f_t(t)$ to the system. Therefore, this value served as a basis for evaluating control performance. $u(t)$ was the amount of control input that the *Human Operator* submitted to the *Controlled System*. This was treated as a measure for control effort. To assess subjective workload, we requested participants to complete a computerized version of the NASA Task Load Index (NASA-TLX) questionnaire [7] after each given visualization.

2 Methods

2.1 Participants

Twelve participants (eight male), were recruited from the participant database of the Max-Planck Institute. They were aged between 21–37 years (mean: 29.1 years) and had normal or corrected-to-normal vision. All were right handed. They gave their written consent before the experiment and were paid 12 Euros per hour.

2.2 Apparatus and Flight Model

The current study was conducted in a fixed-based flight simulator that consisted of a main PC and display cluster. The main PC controlled the experiment and data collection with a customized software based on Matlab Simulink (Mathworks). This PC was connected to a cluster of nine independent visualization PCs, via a local area network and commanded the timing and presentation of the visualization using UDP triggers.

The visualization PCs were connected to a large display that consisted of nine panels (total field-of-view: 105° x 100°). In the *Instrument* condition, two

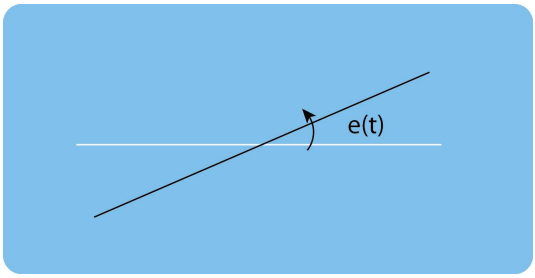


Fig. 2. *Instrument* condition, showing an artificial horizon. The error $e(t)$ is calculated from the reference signal $f_t(t)$ and the roll angle $\varphi(t)$.

lines were rendered on a blue background with Matlab Psychtoolbox, a black line that represented $e(t)$ and a white horizontal line that represented zero error [8,9] (Figure2). The *Outside View* condition used flight simulation software (i.e., FlightGear; [10]) to present a cockpit view of a straight-ahead flight path, through the hinterlands of San Francisco, wherein $e(t)$ resulted in rotations of the cockpit's view frustum and, hence, the entire scene (see Figure 3).

Inputs to the *Controlled System* were submitted via a joystick (Extreme 3D Pro, Logitech) that sampled at 256Hz. This only affected the roll angle of the visualization. The other degrees of freedom of the *Controlled System* were fixed.

A computerized NASA-TLX questionnaire was presented to the participants for the self-reporting of subjective workload via a laptop computer. This rating scale consists of six sub-scales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration) [7].

2.3 Compensatory Tracking Task

In a compensatory tracking task, the participants needs to minimize the error between a target signal and the output of the system. In the current study, a disturbance $f_t(t)$ is continuously introduced into the system, which exclusively perturbs the roll angle of the *Controlled System*. Here, $f_t(t)$ was designed as quasi-random reference signal that consisted of a sum of 10 sine waves. These comprising sine waves were non-harmonically related. The disturbance function had a variance of 1.7 deg^2. More specifically we used the following function [11]:

$$f_t(t) = \sum_{j=1}^{N} A(j) sin(\omega(j) \cdot t + \phi(j)) \quad (1)$$

The amplitude, frequency and phase of the sinusoids are given in Table 1.

2.4 Procedure

Two sessions comprised the full experiment and were conducted on separate days. Two blocks were performed in each session and each block presented one

Fig. 3. Fixed-base flight simulator, consisting of nine panels and a field-of-view of 105° x 100°. Here the *Outside View* is shown. During the experiment, visual disturbances were experienced in the roll-axis around the horizon, that our participants were instructed to compensate for with the provided joystick.

Table 1. Values of the ten non-harmonically related sine waves of the target signal $f_t(t)$. With number of the sine wave j, the amplitude of the j_{th} sine wave equals A_j, the frequency is ω_j and the phase is ϕ_j .

j	A_j in deg	ω_j in rad/s	ϕ_j in rad
1	1.351	0.377	0.145
2	1.007	0.859	0.902
3	0.509	1.759	4.306
4	0.260	2.827	6.127
5	0.157	3.917	5.339
6	0.095	5.466	6.155
7	0.060	7.749	1.503
8	0.043	10.514	1.506
9	0.036	13.132	2.368
10	0.030	17.363	2.086

of the two possible visualizations (*Instrument*, *Outside View*). Three 5 mins trials were presented per block, with 5 mins breaks between them. The order of the blocks was counter-balanced for the visualization condition across sessions and participants.

Each session began with the participant reading and signing a consent form that provided experimental instructions. The computerized NASA-TLX questionnaire was administered after the completion of each block of trials for the

given visualization condition. Altogether, the experiment took 3.5 hours for every participant over the two sessions.

2.5 Data Collection and Analysis

To evaluate the performance, the normalized root mean squared value of the error signal $e(t)$ (nRMSerror) was calculated, as well as the root mean squared value of the control input $u(t)$ (RMSinput). The nRMSerror was normalized with the disturbance that was experimentally introduced to the system. Thus, a nRMSerror that is smaller than a value of 1 would indicate that our participants reduced the disturbance in the system, while a value that was larger than 1 would indicate that the participant introduced additional disturbances to the system. The nRMSerror can be further divided into the mean and variable error as follows:

$$nRMSerror = \sqrt{MeanError^2 + VariableError^2} \tag{2}$$

whereby $MeanError$ is simply defined over all measured time points i as,

$$MeanError = \frac{\sum_{i=1}^{N} e_i}{N} \tag{3}$$

and $VariableError$ as,

$$VariableError = \sqrt{\frac{\sum_{i=1}^{N} (MeanError - e_i)^2}{N}} \tag{4}$$

The mean error represents the distance of the mean of the error distribution from zero (i.e. the target) and the variable error represents the spread of the error distribution [12].

The RMSinput represents the control effort of the participants. A higher RMSinput indicates that the participants submitted more joystick input into the *Controlled System*.

These measures for control performance, control effort and subjective workload were submitted to a paired-sample t-test to test for statistical differences. An alpha-level of 0.05 was adopted as the criterion for significance.

3 Results and Discussion

Figure 4A shows that the *Outside View* visualization resulted in a larger nRMSerror than the *Instrument* visualization ($t(11)=-6.54$, $p < 0.05$). In fact, all of our participants had nRMSerror values that were larger than 1 when the *Outside View* visualization was presented. This means that their efforts to minimize error actually led to additional disturbances in the control system. It is necessary to point out that this was not due to the difficulty of the compensatory

control task per se. When presented with the *Instrument* visualization, all participants were able to achieve nRMSerror values that were lower than 1. This result highlights the critical influence of visualization on control performance.

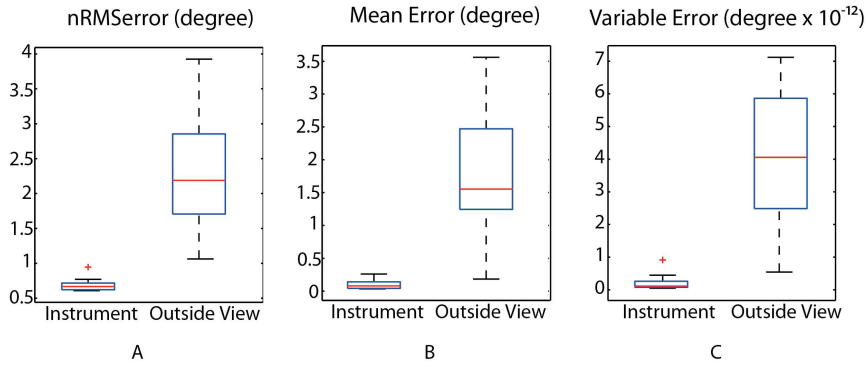


Fig. 4. Box-plots for the measures of nRMSerror (A), mean error (B) and variable error (C) across the condition of Visualization. Each box-plot shows the median, the interquartile range and data range. Outliers are represented as red crosses.

There are several explanations for this large difference in nRMSerror. First, our participants have failed to accurately estimate the desired goal from the *Outside View* visualization. Namely, the ideal attitude. If so, we would expect our participants' error distribution to be shifted away from the zero value, resulting in a bigger bias (e.g. mean error). Next, our participants could have been unable to accurately estimate the error from the *Outside View* visualization. If this was true, we would expect a high variable error. Figures 4B and 4C show that the mean error and the variable error were both larger for the *Outside View* compared to the *Instrument* condition (Mean Error: $t(11)=-5.77$, $p < 0.05$; Variable Error: $t(11)=-6.02$, $p < 0.05$). Therefore, our participants were less certain about the ideal state and were less precise in their control when they were presented with an *Outside View* visualization.

A time trace of the control error for both conditions (Figure 5) shows these two differences between the conditions. In the simple *Instrument* condition (light gray), the control error varied around the target with smaller mean and variable error. In the *Outside View* condition (black) the mean error is shifted over time with larger fluctuations around it.

In addition, the *Instrument* condition resulted in more input activity than the *Outside View* condition ($t(11)=5.59$, $p < 0.05$; see Figure 6). This indicates that the *Instrument* visualization induced our participants to invest more control effort into the task than for the *Outside View* visualization. This could be because error was better perceived from the *Instrument* visualization, resulting in more and better targeted control input. Conversely, participants could have submitted less control input in the *Outside View* visualization because they did not perceive the need for it.

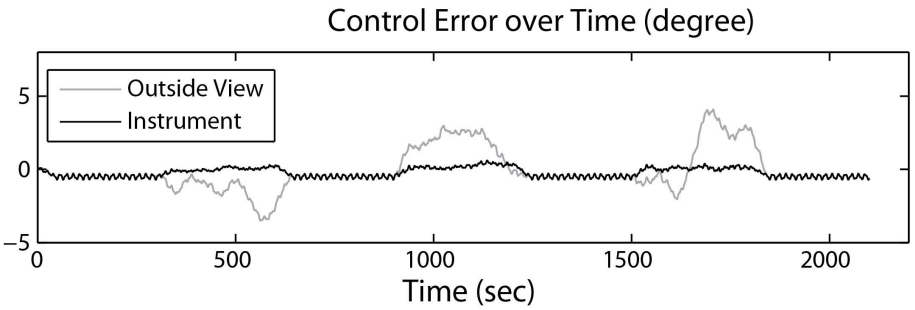


Fig. 5. Control error over time for the *Outside View* (light gray) and *Instrument* (black) condition. The data was filtered using a moving average filter with a window size of 40 seconds. In the *Instrument* condition the error varied around the target while in the *Outside View* condition, the mean of the error distribution is shifted over time.

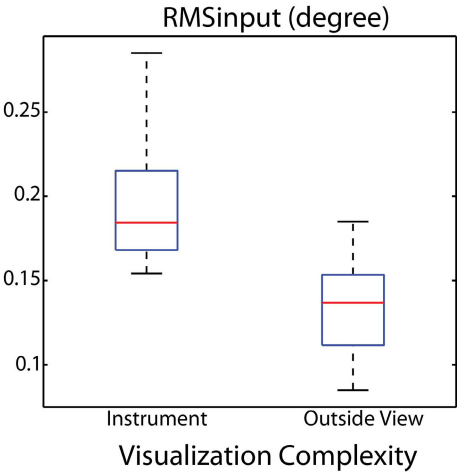


Fig. 6. Box-plots for RMSinput across the condition of Visualization. Each box-plot shows the median, the interquartile range and data range.

Subjective workload, as measured by the NASA-TLX scores, did not differ across the visualization condition ($t(11)=2.07$, $p = 0.07$). This supports our earlier conclusion. Although the participants did not perceive a difference in the difficulty of the same task across the different visualizations, the difference in their ability to accurately perceive their error resulted in very different control performance. The NASA-TLX scores indicate that mental demand (23%), performance (29%) and effort (23%) comprised more than 70% of the perceived workload in our task.

These findings show that control performance is better supported by a visualization that explicitly present the information that is required for the control task. In the current experiment, an explicit representation of the error supported the *Human Operator* in submitted the appropriate control inputs, without increasing his perceived workload. Unfortunately, competence in a complex control task such as piloting an aircraft with many degrees of freedom for a large repertoire of possible maneuvers often depend on multiple sources of information. It may not be feasible to create dedicated instruments for every relevant information channel. In this regard, the outside world might represent a more general and effective source of information than spreading one's visual attention across multiple instruments. This warrants further investigation.

In conclusion, the visualization of the error feedback can result in different levels of performance for two experimental conditions that are equivalent in terms of their difficulty and perceived workload. A simple visualization might lack the qualities of physical realism, but explicitly represents the primary property that is of interest to the human operator. This has the advantage of preventing the occurrence of unintended biases in error perception.

References

1. Browning, A.N., Grossberg, S., Mingolla, E.: A neural model of how the brain computes heading from optic flow in realistic scenes. Cognitive Psychology 59(4), 320–356 (2009)
2. Larish, J.F., Flach, J.M.: Sources of optical information useful for perception of speed of rectilinear self-motion. Journal of Experimental Psychology: Human Perception and Performance 16(2), 295 (1990)
3. Barfield, W., Rosenberg, C., Kraft, C.: The Effects of Visual Cues to Realism and Perceived Impact Point during Final Approach. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 33(2), pp. 115–119 (October 1989)
4. De Maio, J., Rinalducci, E.J., Brooks, R., Brunderman, J.: Visual Cueing Effectiveness: Comparison of Perception and Flying Performance. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 27(11), pp. 928–932 (1983)
5. Zaal, P.M.T., Nieuwenhuizen, F.M., van Paassen, M.M., Mulder, M.: Modeling Human Control of Self-Motion Direction With Optic Flow and Vestibular Motion. IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics 43(2), 544–556 (2012)
6. Reed, M.P., Green, P.A.: Comparison of driving performance on-road and in a low-cost simulator using a concurrent telephone dialling task. Ergonomics 42(8), 1015–1037 (1999)
7. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index). Results of Empirical and Theoretical Research (1988)
8. Brainard, D.H.: The Psychophysics Toolbox. Spatial Vision 10, 433–436 (1997)
9. Pelli, D.G.: The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spatial Vision 10, 437–442 (1997)

10. Perry, A.: The flightgear flight simulator. In: USENIX Annual Technical Conference, Boston, M.A. (2004)
11. Nieuwenhuizen, F.M., Mulder, M., van Paassen, M.M., Bülthoff, H.H.: Influences of Simulator Motion System Characteristics on Pilot Control Behavior. Journal of Guidance, Control, and Dynamics 36(3), 667–676 (2013)
12. Jagacinski, R.J., Flach, J.M.: Control Theory for Humans - Quantitative Approaches to Modeling Performance, ch. 10, pp. 104–109. CRC Press, Mahwah (2002)

Walking Speed in VR Maze while Central Visual Fields Are Restricted with Synchronously Moving Black Circles

Functions of Central Visual Field in Walking through VR Space

Yohsuke Yoshioka[1] and Colin Ellard[2]

[1]Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba, Japan
yoshioka.yohsuke@faculty.chiba-u.jp
[2]University of Waterloo, 200 University Ave W, Waterloo, ON, Canada
cellard@uwaterloo.ca

Abstract. We examined the function of the central visual field by using the newly developed VR system that was consisted with a wide-view HMD and an eye-tracker for restricting an arbitrary area of human visual field. Subjects were asked to walk through short virtual mazes under different visual condition in which 10 or 20 degree of their central visual field was restricted artificially with the system. Results indicated 1) Times for walking through the entire maze under the visual condition with 10 degrees of the central visual field restricted in synchronization were longer than times under the condition in which 10 degrees of the fixed central area of screen were restricted. 2) For walking through the area with two dead ends, walking times under the condition in which 20 degrees of the central visual field were restricted were longer than under the condition in which 10 degrees of the central visual field were restricted.

Keywords: Applied cognitive psychology, Cognitive task analysis, Human Centered Design to reduce through life costs, Human Factors / System Integration, Human Factors certification and regulation, Safety, Simulation.

1 Introduction

The central visual field is consist of the central region of the retina, near the fovea, and has the highest visual acuity of any part of the visual field, and has colour vision. The peripheral visual field is in the large remaining area of the visual field and it deals with low spatial frequency and uncolored vision. The interaction between the two visual fields has a strong influence on human perception and behaviour.

Understanding the functions and relationships of the two visual fields will provide new insight into the organization of spatial perception. It will also draw on a break-through in the theories and techniques for designing more suitable and more realistic VR space.

We attempted to clarify the dynamic functions of the central visual field with a way-finding experiment using a newly developed VR system consisting of Wide-

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 212–220, 2014.

View Head-Mount-Display (HMD) and Eye-Tracker for restricting arbitrary areas of the human visual field. The important data of the results of the experiment will be picked up and shown in this paper.

2 Methods

2.1 System

In the experiment, the subjects were asked to walk through a maze in virtual space under different visual conditions; their central visual field was restricted artificially with the developed VR system.

The system consisted of a wide-view HMD (Nvis: nVisor SX111), an eye-tracker (Arrington Research: Binocular Eye-tracking system), and a position tracking systems (WorldViz: PPT optical tracker). The position of the subject in the real experimental room was tracked with the position tracking system consisting of eight high-resolution cameras arranged in the four corners of the experimental room.

The VR space displayed on the HMD was linked with the calculated position of the subject in the real experimental room so that the subject could walk and look around the VR space on foot. Wide-View HMD allows subjects to view 111 diagonal degrees of the visual field on one screen, which covers much of the peripheral visual field of humans.

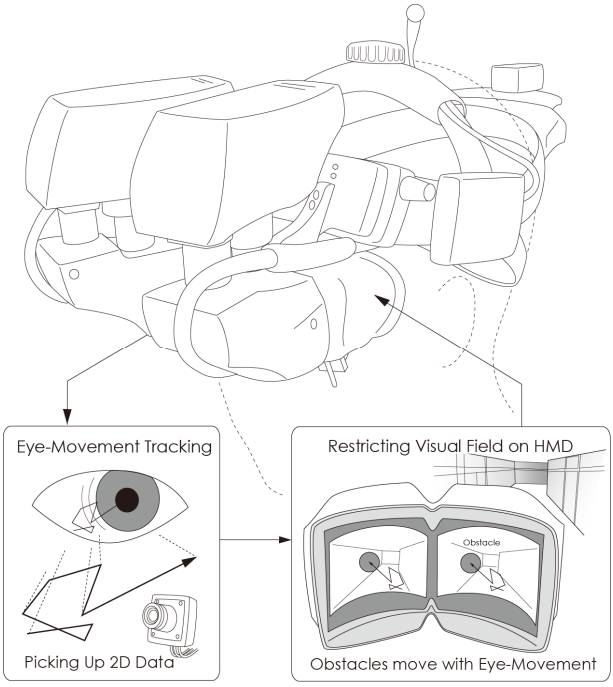


Fig. 1. Arrangement of HMD and Eye-tracker in Experimental system

The most specific feature of the system is that it also has a binocular eye-tracking system in the HMD (Fig. 1). Small precision cameras and two pieces of mirror were combined and installed into the clearances between the HMD screen and eyeball in order to record the subjects' eye movement. The recorded eye movement was sent directly to the workstation connected with the HMD, and the precise position of each subject's fixation could be calculated with software simultaneously.

We used Vizard 4.0 (World Viz) as VR software for describing the virtual world on the screen of the HMD. By sending the real-time fixation position to the VR software, the developed system could modify the display in concert with the subject's fixation pattern on the VR world. In this study, for example, we displayed small black shields in the VR world, which were controlled in synchronization with the eye movement in real time for restricting the view to a specific area of the visual field. The dynamic relationships between some parts of the visual field could be clarified with the way-finding experiment using the system.

2.2 Mazes

Figure 2 shows the four patterns of the short virtual mazes that the subjects walked through in this series of the experiment. Every maze was the same size and arranged within an area of three meters by five meters with three-meter-high virtual walls. Each of them had a three-way intersection at the starting point in order to examine the behaviour of the subjects in a situation where they had to choose the correct way to the goal from these three similar looking ways under the restricted visual condition.

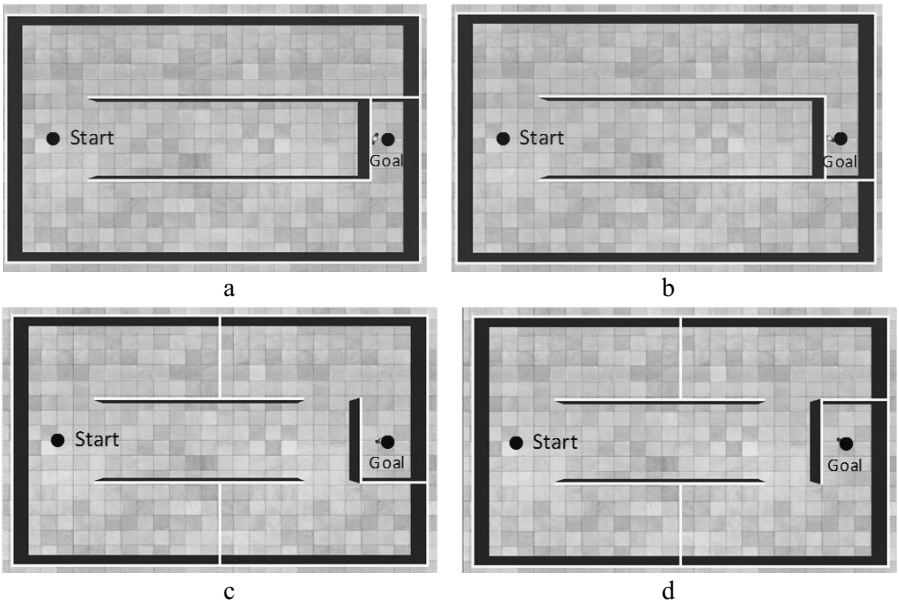


Fig. 2. The Patterns of the Virtual Mazes

Maze pattern (b) was arranged as the reverse of maze pattern (a). These two patterns were alternated as the first choice of each subject. If the subject first chose the left direction from the starting point, the route would reach a dead end so that a piece of the wall would come appear on the left side of the goal.

Because of the interactive system, the abilities of the subjects in some different restricted visual conditions to return from the deeper dead-end could be tested in every trial.

Maze patterns (c) and (d) were designed in the same geometrical relation. At the starting point, the subject had to choose the middle of the three forks as the only correct route to the goal. After that, they immediately came to an interactive junction just before the goal. If they chose the left option, the route became a dead end; a piece of the wall appeared on the left side of the goal.

The abilities of the subjects to subsequently choose the direction they had not yet been to could be tested in every trial.

2.3 Visual Conditions

Each subject walked through the short virtual mazes from the starting point to the goal 16 times in the 4 different maze patterns above and under the following 4 visual conditions (Fig. 3):

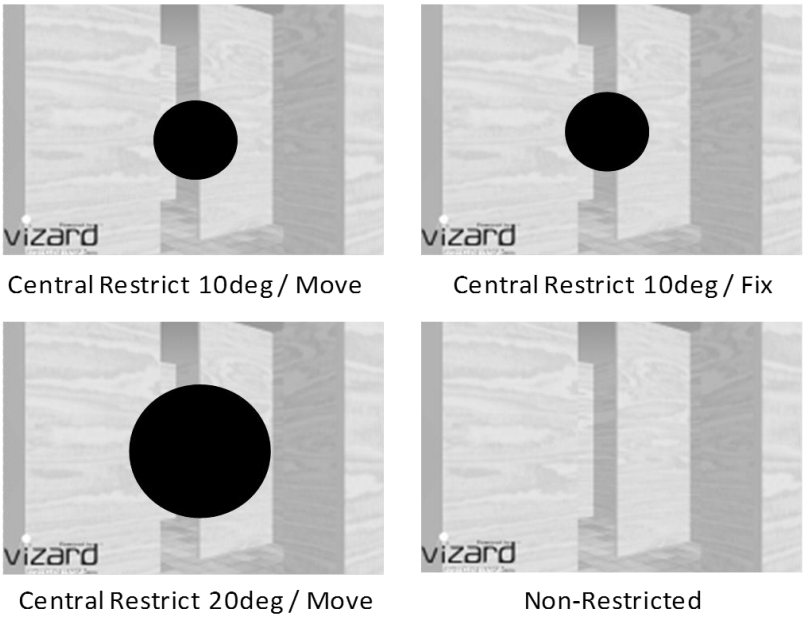


Fig. 3. The Four Visual Conditions

- **Central Restrict 10 deg/Move**
  The central visual fields of the subjects were restricted by virtual black circles moving synchronously with the eye movement. The size of the circles could cover the area of 10 degrees around of the fixation points.
- **Central Restrict 10 deg/Fix**
  Black circles of the same size as in the "Central Restrict 10 deg/Move" condition were displayed at the center of the screen of the HMD, but they were fixed on the center and not moving with eye movement. The effects of the eye-following movement of the circle could be examined by the comparison of the subjects' behaviour under the two conditions, moving and fixed.
- **Central Restrict 20 deg/Move**
  The black circles were moving with the eye movement, and the sizes of the black circles were larger and covered the area 20 degrees around of the fixation points. The effects of the size of the restriction could be found by comparing the 10-degree and 20-degree conditions.
- **Non-Restricted**
  There were no restrictions on the subjects' visual fields. As the basic condition, the subjects simply walked through the virtual mazes with their normal vision.

2.4 Subjects and Informed Concent

A total of 20 college students took a part of the experiment as healthy subjects. This experiment was approved by the University of Waterloo ethical committee. Written informed consents were obtained from all subjects for publication of this case report and accompanying images.

3 Results

3.1 Times for Walking through the Entire Maze

This paper particularly deals with the time it took to walk through mazes (a) and (b) as the main results of the experiments.

This paper did not make any mention about the data taken from the maze(c) and (d). However, because of the insertion by those two types of maze, the middle route in front of the starting point would be not always a dead-end. The subjects were required to take a time for confirming the detailed shape of the middle route, even if at the trials in mazes (a) and (b).

An ANOVA showed significant differences in the mean between the four visual conditions in mazes (a) and (b). Fig.4 shows the average times for all subjects walking through the mazes under the four visual conditions. The P values as the differences in the mean values were assessed with a Bonferroni multiple comparison procedure.

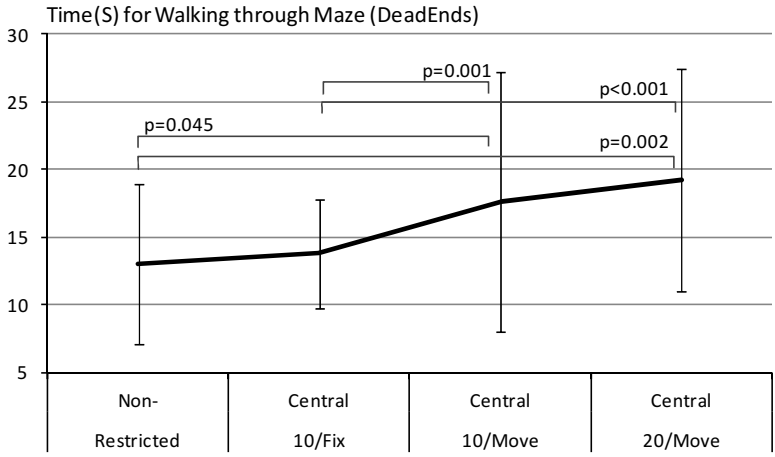


Fig. 4. Times for Walking through Mazes

There are two important significant differences between the mean values in the figure. The first is the difference between the walking times under "Central Restrict 10 degrees/Move" and "Central Restrict 10 degrees/Fix." The only difference between those two visual conditions was that the black circles were moving with eye movement under the Move condition but not under the Fix condition.

The second significant difference is that the walking times under "Central Restrict 10 deg/Move" and "Central Restrict 20 deg/Move" were longer than under the "Non-Restricted" condition.

Since the time for walking through the maze under "Central Restrict 10 deg/Fix" was not longer than under the "Non-Restricted" condition, results suggest that it was not so difficult to walk through the maze under the fixed restriction. On the other hand, it became significantly more difficult when the same sized restriction moved with eye movement.

3.2 Times for Walking through the Each Area of the Maze

In order to analyze the data more closely, we calculated separately the time for walking in the three divided areas showed in Fig. 5.

- **Area Dead-Ends**
  Including the starting point and the two deeper dead-ends. The subjects were required to recognize the detailed shape of the dead-end, especially the connectional relations between the walls, and they were also required to come back to the starting point after that.
- **Area Non Dead-End**
  Including the correct route to the goal. The subjects finally could walk through this area as the third selected route after coming back from the two complicated deeper dead-ends.

- **Area Goal**
  There is a floating icon behind the corner of the goal. The subjects were required to find and touch it out for finishing each trial.

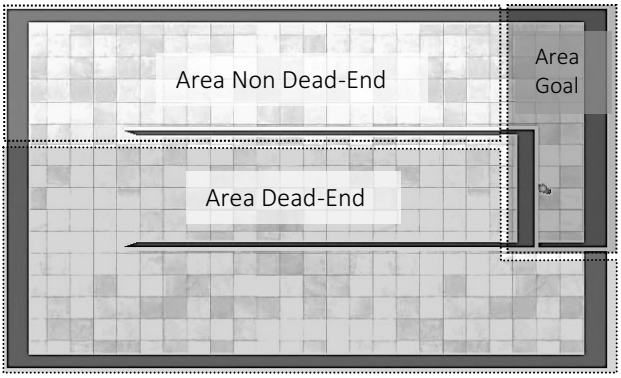


Fig. 5. Divided Area of Maze

The significant differences we found in the time for walking through the entire maze are found again in Fig. 6, the average walking time of all subjects in "Area Dead Ends."

It is noteworthy that there is another significant difference between the two moving conditions. This result suggests that it was more difficult under "Central Restrict 20 deg/Move" than under "Central Restrict 10 deg/Move" to walk in "Area Dead End" where the subjects were required to recognize the detailed shape of the dead-end as quickly as possible degrees.

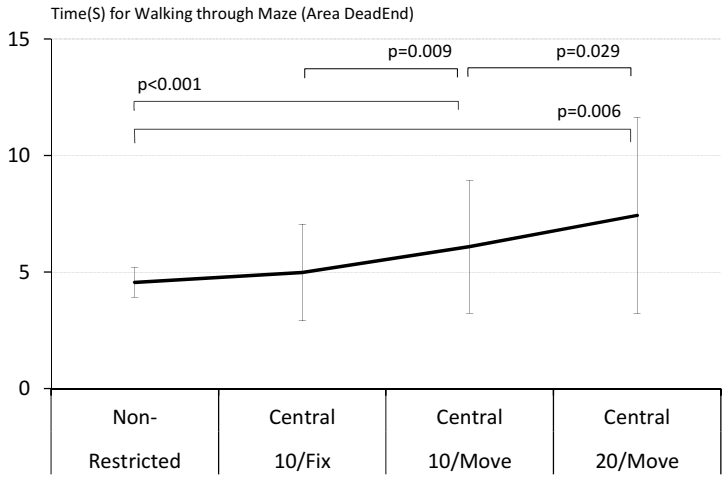


Fig. 6. Times for Walking through "Area Dead Ends"

4 Conclusions

The following can be made from the results of the experiments:

- Times for walking through the entire maze under the visual condition with 10 degrees of the central visual field restricted in synchronization with eye movement were longer than times under the condition in which 10 degrees of the fixed central area of screen were restricted.
- For walking through the divided area with two dead ends, walking times under the condition in which 20 degrees of the central visual field were restricted were longer than under the condition in which 10 degrees of the central visual field were restricted.

The findings in this experiment demonstrate the use of novel technology to help understand the processes that take place during human way-finding. Our finding that dynamic occlusion of the central visual field as the subjects walked through the maze produced the most pronounced changes in way-finding behaviour demonstrates that this part of the visual system makes a significant contribution to route selection.

Acknowledgements. We would like to express our deepest gratitude to Deltcho Valtchanov, Kevin Barton and Danniel Varona-Marin who provided carefully considered feedback and valuable comments. This work was supported by Japan Society for the Promotion of Science Grant-in-Aid for Young Scientists (A) No.22686056, and The Researcher Exchange Fellowship of The Natural Sciences and Engineering Research Council of Canada.

References

1. Anstis, S.M.: A chart demonstrarin variations in acuity with retinal position. Vision Research 14, 589–592 (1974)
2. Brown, B.: Resolution thresholds for moving target at the fovea and in the peripheral retina. Vision Research 12, 293–304 (1972)
3. Committee on Colorimetry, Optical Society of America (1963)
4. Dolezal, H.: Living in a world transformed: perceptual and performatory adaptation to visual distortion. Academic Press Proceedings Series 26, 161–184 (1982)
5. Ikeda, M., Saida, S.: Span of recognition in reading. Vision Research 18(10), 83–88 (1978)
6. Ikeda, M., Takeuchi, T.: Influence of foveal load on the functional visual field. Perception & Psycholophysis 18, 255–260 (1975)
7. Johansson, G.: "Visual Perception of locomotion elicited and controlled by a bright spot moving in the periphery of the visual field" Report #210 Department of Psychology Uppsala University (1977)
8. Kundel, H.L., Nodine, C.F., Toto, L.: Eye movements and the detection of lung tumors in chest images. In: Gale, A.G., Johnson, F. (eds.) Theoretical and Applied Aspects of Eye Movement Research, pp. 297–304. Elsevier Science, New York (1984)

9. Kuroiwa, M., Okazaki, S., Yoshioka, Y.: Comparison of visual behaviors within a normal visual field and a restricted visual field: navigating a corridor and staircase. The Japanese Journal of Ergonomics 37(1), 29–40 (2001)
10. Livingstone, M., Hubel, D.: Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. Science 240, 740–749 (1988)
11. Low, F.N.: The peripheral visual acuity of 100 subjects. American Journal of Psychology 146, 573–584 (1946)
12. Mateef, S., Gourevich, A.: Peripheral vision and perceived visual direction. Biological Cybernetics 49, 111–118 (1983)
13. McConkie, G., Zola, D., Wolverton, G.: Time course of visual information utilization during fixations in reading. Journal of Experimental Psychology. Human Perception and Performance 10, 75–89 (1984)
14. Maunsell, J.H.R., Newsome, W.T.: Visual processing in monkey extrastriate cortex. Annual Review of Neuroscience 10, 363–401 (1987)
15. Osaka, N.: Peripheral lower visual field: a neglected factor? Behav. & Brain Sciences 13, 555 (1990)
16. Paap, K.R., Newsome, S.L., McDonald, J.E., Schvaneveldt, R.W.: "An activation-verification model for letter and word recognition" The word-superiority effect. Psychological Review 89, 573–594 (1982)
17. Polyak, S.L.: The retina. University of Chicago Press, Chicago (1941)
18. Previc, F.: Functional Specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. Behav. & Brain Sciences 13, 519–575 (1990)
19. Saida, S., Ikeda, M.: Useful Visual Field Size for Pattern Perception. Perception & Psychophysics 25(2), 119–125 (1979)
20. Skrandies, W.: Human contrast sensitivity: Regional retinal differences. Human Neurobiology 4, 97–99 (1985)
21. Ungerleider, L.G., Minshkin, M.: Two cortical visual systems. In: Ingle, D.J., Goodale, M.A. (eds.) Analysis of Visual Behavior. MIT Press (1982)
22. Warren, R.: The perception of ego motion. Journal of Experimental Psychology Human Perception and Performance 2, 448–456 (1976)
23. Wertheim, T.: Über die indirekte Sehschärfe. Zeitschrift für Psychologie und Physiologie der Sinnesorgane (Leipzig) 7, 172–187 (1894)
24. Woodhouse, J.M., Barlow, H.B.: Spatial and temporal resolution and analysis. In: Barlow, H.B., Mollon, D.J. (eds.) The Senses. Cambridge University Press, Cambridge (1982)

Cognitive Issues in Interaction and User Experience

Towards a Context Model for Human-Centered Design of Contextual Data Entry Systems in Healthcare Domain

Maxime Baas[1], Stéphanie Bernonville[2], Nathalie Bricon-Souf[3], Sylvain Hassler[4], Christophe Kolski[1], and Guy Andre Boy[5]

[1] UVHC, LAMIH, F-59313 Valenciennes, Univ Lille Nord de France, F-59000 Lille, CNRS, UMR 8201, F-59313 Valenciennes, France
maxime.baas@free.fr
[2] EA 2694, Univ Lille Nord de France, Service Information et Archives Médicales, CHRU Lille, F-59000, Lille, France
[3] Université de Toulouse, UPS-IRIT - plateforme e-santé, ISIS Rue Oules F-81100 Castres, France
[4] Inserm CIC-IT, Univ Lille Nord de France, CHRU de Lille, UDSL EA 2694, F-59000, Lille, France
[5] Human Centered Design Institute, Florida Institute of Technology, 150 West University Blvd, Melbourne, Florida 32901, USA

Abstract. Data entry by physicians is a critical aspect in the health care domain, in which errors may lead to severe consequences for patients. This paper describes and discusses these aspects to support human-centered design of appropriate human-computer interaction technology. The following issues will be addressed, including system aim, users' profiles, interaction devices and environment of use, to cite the most important. Our work is based on a literature survey, questionnaires, and an active participatory design process conducted with healthcare professionals. Since the crucial factor is context of use, we elicited several relevant contextual attributes that enabled us to create and incrementally upgrade a context model. This conceptual model is intended to support a scenario-based design approach of future data entry systems. A few scenarios are provided.

Keywords: Data entry, Context, Health Care, Human-Centered Design, User, Input Device, Environment.

1 Introduction

One of today's major challenges is to provide healthcare professionals with an easy-to-use data entry system (e.g., for prescription data entry, medical records). Our main goals are to reduce the number of input errors in normal, abnormal and emergency situations [1][2], to structure data entry and to reuse medical data. Various kinds of users will be considered, including physicians, nurses and medical assistants. All of them will interact with various kinds of interaction devices using their own background and practice [3]. These devices will enable healthcare personnel to create, modify or remove information using a graphical interface. Many data entry systems can be candidates to support these tasks [4].

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 223–233, 2014.

For Boy [5], "Context can be defined along several attributes, such as task […], space […] and time […]." For Dey [6], "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." With these definitions in mind, we analyzed the literature on context, several input systems and acquired data from questionnaires fulfilled by healthcare professionals with respect to different work situations that require data entry. A thorough analysis of the questionnaire answers led to the definition of relevant context attributes, which in turn contributed to the creation of a context model. This model will support the design for future data entry systems.

We first present the importance of taking into account context in life-critical systems design. Then, our method that consists in identifying context attributes in health care is explained. Consequently, a first version of the context model is proposed. Scenarios showing how it can be used are presented. Finally, the conclusion presents further work.

2 Importance of Context in Design of Life-Critical Systemas

An artifact is a concrete materialization of a concept. By definition, it is then limited by the context of definition of the concept being materialized. Consequently, when this artifact is used or operationalized in a different context, it may not be appropriate or adapted. We usually talk about rigidity. Automation is a good example of rigidifying work practices, and in some situations it may lead to surprises. Human-centered automation was developed to include context features into automata. Context can be viewed from various perspectives. First, context can be considered as entities that disambiguate a situation, an event or an affordance, in order to guide action. In other words, context provides pragmatics to human-system integration. Second, context can be considered as an environmental model that provides meaning to a concept. It provides time, space and other conditional factors that refine affordances of a concept or an artifact.

In this paper, context refers to human-computer interaction practices in healthcare, and more specifically data entry systems. It encapsulates factors related to the types of artifacts (e.g., interaction devices), users, tasks, organizations, and (environmental) situations. We refer to the AUTOS pyramid [5][7][8]. AUTOS provides a framework to capture contextual processes, factors and attributes. It is successfully used in life-critical systems such as aeronautics, space, automobile and telecommunication. We will then describe contextual factors with respect to these high levels AUTOS classes.

Most physicians enter medical data using paper and pencil. The main problem is the rigidity of this medium, since when secretaries are available, they need to re-enter part or totality of data into computers to enable information transfer among various actors, including patients, other physicians and administrative services. Our digital world imposes more effective ways of transferring data among actors. This is why it is urgent to better understand how health care data has to be entered into digital systems. It obviously depends on context. It depends on the type of input device (e.g., pen-based, keyboard, voice recognition), user (e.g., physician, medical assistant or nurse), task (e.g., medical prescription, request for medical analysis or investigation), organization (e.g., large hospital or private practice), and situation (e.g., prior to an operation, emergency or regular examination).

3 Method Used for the Identification of Context Attributes

We used the following method to identify context attributes in work situations involving physicians: (a) literature review and study of existing healthcare data entry systems; (b) use of paper and electronic questionnaires addressed to heath care professionals. Results were analyzed and led to issues that were evaluated by heathcare professionals, figure 1.

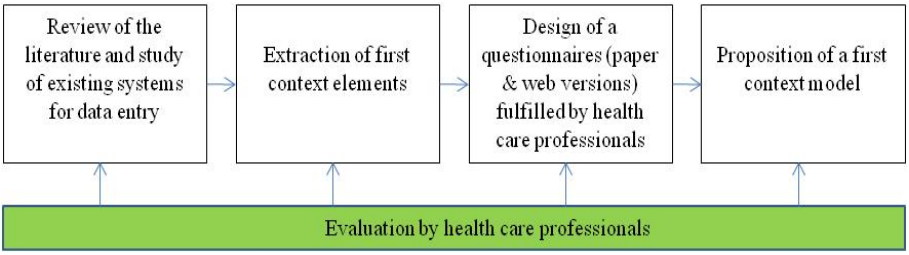


Fig. 1. Method used

3.1 Literature Review and Study of Existing Data Entry Systems

The literature review first enabled the identification of general and domain-specific data entry descriptors (keywords). Keywords were identified during brainstorming sessions involving HCI specialists, psychologists and medical informatics specialists. 133 keywords was created and classified into 4 categories (human factors, data entry, medical, other words). We have combined each keyword with each category (our combination is: (Human Factors) AND (Data entry) AND (Medical) AND (Other word). Such combinations were used to search papers into 3 databases (ScienceDirect, ACM and pubmed). Finally, we have selected 134 papers. The systems or data entry approaches described in these papers have been analyzed.

In this section, six representative data entry systems will be described. They were chosen for their complementary interaction modalities. They are described below. The first three systems are used by healthcare professionals. The others are intended for use by disabled people.

The first medical data entry system is ISME (Icon-based System for Managing Emergencies) [9]. It allows communication between different first-aid teams during a crisis (accident, fire, gas leak, etc.). In order to provide an effective data entry system, which can be understood by all first-aid teams, ISME uses an iconic language. This iconic language is created from different crisis scenarios. However, it was not user-centered designed and does not follow the ISO recommendations for the creation of icons (norm IEC 80416-1: 2001) [10].

The gestural data entry system with linguistic recognition suggested in [11] saves the data using an electronic pen. This system is intended for use in hospitals. The data entry is done on a paper covered with dots, which make it possible to calculate the position of pen on the paper. When the pen is connected to a computer, the recorded

positions are grouped to transcribe data in a computerized format. In this system, writing recognition is limited because the healthcare professional can only fill in predefined boxes. If the user needs to type complementary information, this data will not be recognized when the document is saved. It will only be "saved" on the paper document and not in the system database.

For medical data entry in hospitals, Alapetite [12] uses voice data entry for anesthetists in the operating theatre. In this context, the choice of a vocal data entry system allows the anesthetist to input information without touching the device (here a computer). To increase data capture quality, Alapetite classified the different noisy elements in an operating theatre, with the aim of reducing the interference of background noise. However, the choice of a dictionary (the list of words that the linguistic system can recognize) is not described in the paper. If the dictionary is well built, the vocal recognition system will be more efficient and more accurate. The choice and structure of the words in the dictionary are very important. Indeed, a dictionary with few words gives a more accurate degree of recognition but increases the risk of mistakes, mainly if the user uses a word that is not included in the dictionary. The more complete a dictionary is, the more the probability of the system finding the word is increased, but the risk of conflicts related to linguistic recognition is also increased [13]

Braille is a language used by blind and visually impaired people to read a document or a book using fingers [14]. For people with a serious visual impairment, there are many keyboards designed for them specifically. To understand the use of these keyboards, we need to know what a Braille character is like. A Braille character is composed of six dots divided into two columns of three dots. Most Braille keyboards have seven buttons, six buttons for the dots and one button to validate the character. Familiarity with keys (for physical keyboards) and buttons (for virtual keyboards) requires an adaption period to locate positions correctly. In virtual versions, button localization is more difficult. This is due to the fact that buttons do not have a haptic return movement that helps localization.

The K-Thôt virtual keyboard [15] is an input system designed for a user suffering from cerebral palsy. The keyboard is designed to reduce user movements. In order to decrease both number and amplitude of mouse movements, each button has two actions, one with a left click and the other with a right click. At the moment, this system does not allow optimal input. The addition of input assistance with a context recognition system would increase input speed when using K-Thôt.

The most widely used input assistance tools for disabled users usually have a word prediction system. For example, Sibylle [16] is a predictive keyboard that helps the user to type more rapidly than with a classical keyboard by suggesting most likely words and letters during typing. The dictionary used in Sybille is based on a journalistic corpus. If the user is writing in something close to a journalistic context, the assistance system gives very good results. However, when writing in another context (using specific jargon, for example), the prediction system is less efficient.

These six representative data entry systems are usable in very different contexts. Each of them provides us with different possible context attributes. These attributes helped us to propose the following questionnaires.

3.2 Proposition of Questionnaires Fulfilled by Healthcare Professionals

From the review of the literature, we have identified a first set of context attributes and created a first context model. But, this first context model was very generic. We have converted this model into a questionnaire (existing in two versions: paper and web site). Each element of the context model has been translated into a question. The aims of the questionnaire were (1) to receive descriptions by health care professionals of contexts about general practice and specialized ones (e.g. Anaesthetics, Geriatrics, Neurology), (2) from such descriptions, to enrich the first context model and progressively transform it into a model usable in health care domain.

The questionnaire was composed of 5 parts. The first part is an identification page. The second is about the context from a user point of view. The third part concerns the concern from the platform point of view. And the last part is about environmental aspects of the context.

The questionnaire has been sent to physicians. We have obtained 22 answers with 12 answers with double profiles (general practice and specialist in the table 1). Examples of relevant environmental aspects are visible in Table 1.

4 Proposition of a Medical Context Model

Using data from the literature review and the answers to the questionnaires, we created a medical context model for data entry (see Fig. 2).

Table 1. Examples of attributes identified from the answers to the questionnaire on environmental context (extract)

	Number of Answers	Mode of answers	Total	Profile	
				General Practice	Specialist
Noise level	21	25 dB	6	4	5
		50 dB	6	5	5
		60 dB	12	12	12
		70 dB	5	5	5
		100 dB	5	5	0
Brightness level	21	Low	4	4	4
		Moderate	12	12	12
		High	3	3	0
Available networks	22	No network/ Fax	1	0	1
		3G	10	10	10
		Wireless	13	12	11
		Ethernet	14	12	12

Fig. 2. Context model for data entry

4.1 Description

We used the definition found in Coutaz and Calvary [17] to categorize each context attribute. Three categories were initially used: the user (the type of person likely to use the system), the device (software and hardware elements of the system) and the environment (where the system is to be used). We add two categories: the task and the organization to comply to the AUTOS pyramid (see §2).

In the user category, it is necessary to define knowledge, the physical state and the task of the user during the use of the system. Professional knowledge, associated to computing knowledge will enable users to guide data entry and choose a couple [input device (e.g., computer, tablet, interactive table) – data entry mode (e.g., keys, gestural or voice)] suitable for the user. Awareness of the physical state and possible handicaps of the user help creating a more usuable and comfortable data entry system. User's knowledge and vocabulary are determined by his/her job or field of activity, e.g., abbreviations and technical words. Meaning of these terms is context-dependent. In the healthcare domain for example, systems must use healthcare technical terms, vocabulary and major abbreviations, which can be easily recognized and reused.

The device category defines software and hardware tools. They can provide input assistance aimed at completing data entries. The characteristics of the device and particularly the choice of the screen (e.g., size, resolution, touch-sensitive) can be important attributes for data entry. Indeed, the smaller the device, the less information can be displayed, and the harder it is to select the target with a pointing device (mouse pointer, finger or tangible object) [18][19]. Like for users, the device must know and recognize main technical terms. Before being able to recognize them, it is necessary to define how these terms will be written and displayed, for example as text, icon, handwriting or vocal recording.

The environment category represents where the system is used. A noisy environment will reduce the efficiency of a vocal system. Like the ISME [3], the environment can be dynamic. For example, a fire explosion may lead to toxic smoke and consequently to many injured people. An efficient input assistance system should be able to propose icons for "toxic smoke" and "injured people" once the "explosion" term is generated. Several possible solutions can be proposed such as an assistance system that is able to anticipate future situations and thus the user's data entry, or an Internet network that would share data or intelligently capture information.

The task category represents the goal(s) of the human-machine system. For identifying a task (or sub-task), one need to know: its name and aim, number of users, who is the user(s) and if the users are mobile during the task performance.

The organization category includes all users, all platforms and all environments interacting with the user who performs the task. In the medical domain, hospitals or private practice can be considered.

4.2 Specification of a Contextual Data Entry System

With the context model mentioned above, system context awareness is likely to help physicians during data entry. We propose a data entry distributed system to facilitate

data entry. This system is composed of three parts: a context model, a context engine, and healthcare web services.

The role of the context model is to help capturing current context of data entry. In the medical domain, global context is defined by at least two actors: the physician with his or her medical specialties and the patient with or her age, gender and medical history. The platform supports information management/saving/capture from hardware and software points of view; medical information may use different standardized terminologies, such as the SNOMED CT (http://www.ihtsdo.org/snomed-ct) code, or Vidal Recos (www.vidal.fr)… The environmental context allows considering different types of data, such as work location (e.g., hospital, physician's office) or presence of internet networks.

The context engine aims at taking into account the current context and choosing the most useful web services. It uses a catalog of web services, characterized by their own identity cards, i.e., inputs/outputs and use contexts. It provides the most adapted web services, and sends appropriate information to these web services and waits for their answers (0 to n answers). When answers are redundant (presence of doubloons), doubloons are removed.

Medical resources are provided through web services. They can be a terminology (ex: SNOMED CT), a best medical practice or a dictionary (synonyms, abbreviations, acronyms).

One example of web service is an iconic language. We propose to use an iconic language such as VCM (*Visualisation des Connaissances Médicales*; Visualization of medical knowledge) [20][21]) to facilitate reading and data entry ; its simplifies reading and automatically identify medical data, using icons. The use of this language allows healthcare professional to have a visual and easy-to-read summary of medical data. Furthermore, VCM offers a synthetic view system called *Mr VCM*. Its objective of is to represent the different parts of the human body in the form of a silhouette with VCM icons, for example heart diseases are represented by a red icon with a white heart shaped design inside. *Mr VCM* can thus be used as a synthesis system.

5 Example of Scenario

During the bedroom tour in a hospital ward, the hospital physician sees his/her patients, accompanied by with one or several nurses. For each patient, the physician has a medical record with the latest information and developments in the patient's state of health (result of analyses, nurses' reports, etc.). The ward has its own secure wifi access and the hospital saves the information on several servers located in various places.

To focus on the context elements, we class them in one of five context categories (according with the AUTOS model):

1. Artifact (interaction device): as the user has to perform a task in mobility, the device has to be mobile too. The use of tactile tablet may be recommended.
2. User: the user (physician) has medical knowledge and more specifically mainly in a field linked to the ward. Each patient is supposed to have medical records.

3. Task: the user has to understand the previous and current states of each patient, to make decisions and add supplementary information (data entry).
4. Organization: the user (physician) interacts with several nurses and patients.
5. (Environmental) situation: a wireless network will help to use the tactile tablet. The physician is mobile from a room to another.

After the save of the context attributes (automatically or through different human-computer and human-human interactions), the context engine sends the concerned context model to dictionary of health web services (Fig. 4). The dictionary will propose the web services the most adapted with the context model. When the dictionary chooses the web services, the dictionary sends the model context to web services. Each web service receives the model context and proposes the best answer. Each answer is sent to the context engine.

As the physician consults and inputs data from room to room, it is relevant for him or her to have a mobile device such as a smartphone or tactile tablet. The device contains medical records for the patients on the ward. One supposes that the VCM (*Visualisation des Connaissances Médicales*; Visualization of Medical Knowledge) iconic language is used. A model illustrating these principles is shown on Figure 3.

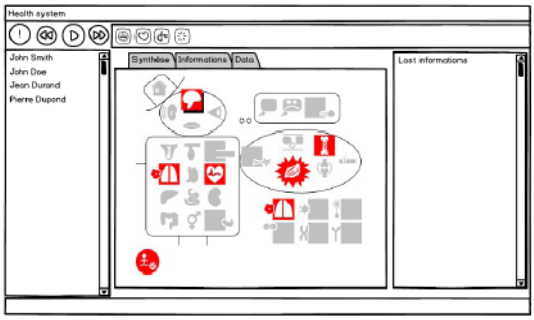


Fig. 3. Mock-up of the context sensitive system proposed for our scenario, using *Mr VCM*

At the center, the physician can see a synthesis on the patient's current state of health and a second tab gives access to patient's full medical records. On the right, the physician can see a list of the latest information about the patient, representing the evolving context linked to his/her health. On the left, the physician has a list of patients he/she needs to see during the round. In the environmental context, the reader system is present above the patient list. Indeed, with the next button, the physician will be able to go on to the next patient, without having to search on the list.

The play / pause button is used to stop the consultation in the event of an emergency on the ward. This button is used to stop the round and display data on the emergency in question. Once the emergency is over, the physician can press the button again to resume the round of the rooms where it had stopped.

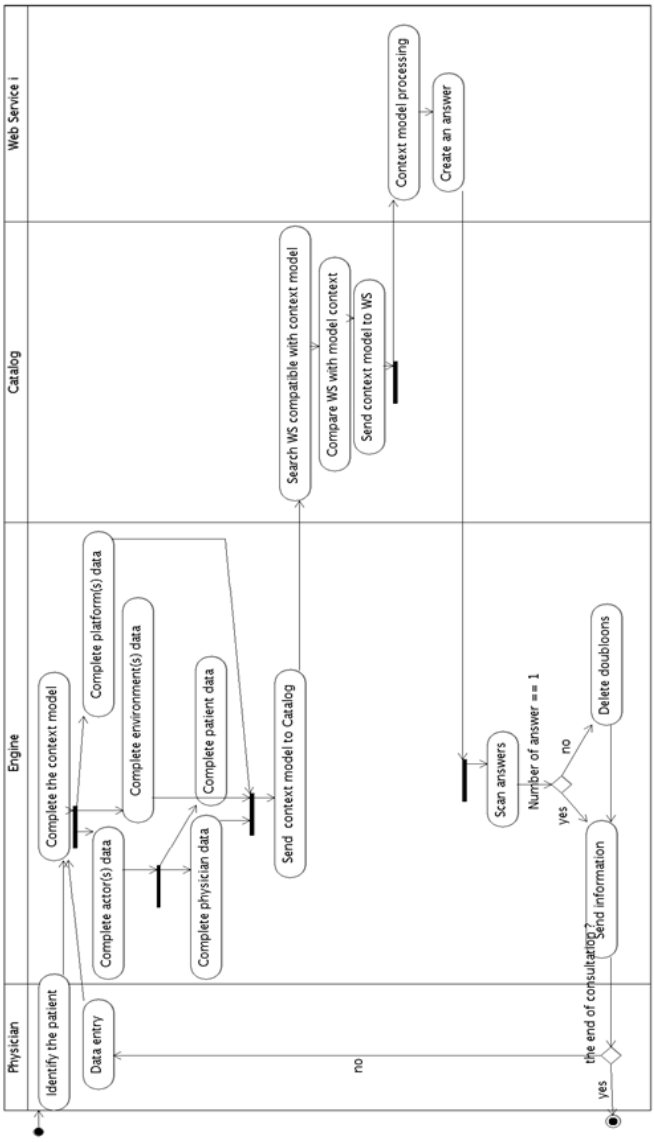


Fig. 4. Context engine logic

This case study shows that taking the context into account in data entry can be a complicated task. Each element of the context may increase or decrease the quality of the entry. This is why the study of context in the design of healthcare systems should be considered with great care and offers many opportunities.

6 Conclusion

A context model for use in the field of health care has been proposed. This model was developed from (1) a critical analysis of existing systems found in the literature and (2) the analysis of questionnaires addressed to heath care professionals. Using the AUTOS pyramid framework, we derived a set of context attributes: the organization; the future users of the system; the task to perform; the device which includes the software and hardware that contribute to the interaction with the system; and the environment in which the system will work. The aim of this model is to assist in the consideration of context in the design of systems for data entry. The model was produced as part of a national project aiming to facilitate medical data entry for health care professionals. One of the perspectives of this model is to be completed over time to create a model that can be used whatever the field of use. In the medical field, context enables to provide systems that are familiar with the field and are therefore intuitive, in order to better help physicians. To allow detection of contextual elements in a project, the model will be adapted in the form of a questionnaire intended for future users of the system. In the future, the context model will be part of system design in the medical field and then in any other field. Here, the objective is to provide a system or context mapping aid in a project to improve future systems.

Acknowledgements. The authors wish to thank the ANR and partners of ANR TecSan 2011 SIFaDo project (n° ANR-11-TECS- 0014).

References

1. Bricon-Souf, N., Newman, C.R.: Context awareness in health care: A review. Int. J. Med. Inf. 76(1), 2–12 (2007)
2. Schilit, B., Theimer, M.: Disseminating active map information to mobile hosts. Presented at the IEEE Network 8(5), 22–32 (1994)
3. Stephanidis, C.: The universal access handbook. In: Stephanidis, C. (ed.). CRC Press, Boca Raton (2009)
4. Martin, B., Pecci, I.: État de l'art des claviers physiques et logiciels pour la saisie de texte. Revue D'Interaction Homme-Machine 8(2), 147–205 (2007)
5. Boy, G.A.: Cognitive Function Analysis. Ablex Publishing Corporation (1998)
6. 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, London, UK, pp. 304–307 (1999)
7. Boy, G.A.: Orchestrating Human-Centered Design. Springer (2012)
8. Boy, G.A.: The Handbook of Human-machine Interaction: A Human-centered Design Approach. Ashgate (2011)
9. Fitrianie, S., Rothkrantz, L.J.M.: Communication in Crisis Situations Using Icon Language. In: ICME, IEEE International Conference on Multimedia and Expo (ICME 2005), pp. 1370–1373 (2005)

10. Lamy, J.-B.: Conception et évaluation de méthodes de visualisation des connaissances médicales: mise au point d'un langage graphique et application aux connaissances sur le médicament. Ph.D. Thesis, Université Paris 6 (2006)
11. Estellat, C., Tubach, F., Costa, Y., Hoffmann, I., Mantz, J., Ravaud, P.: Data capture by digital pen in clinical trials: A qualitative and quantitative study. Contemp. Clin. Trials 29(3), 314–323 (2008)
12. Alapetite, A.: Impact of noise and other factors on speech recognition in anaesthesia. Int. J. Med. Inf. 77(1), 68–77 (2008)
13. Gong, J., Tarasewich, P., Hafner, C.D., Mackenzie, S.I.: Improving dictionary-based disambiguation text entry method accuracy. In: CHI EA 2007: CHI 2007 Extended Abstracts on Human Factors in Computing Systems, pp. 2387–2392 (2007)
14. Southern, C., Clawson, J., Frey, B., Abowd, G., Romero, M.: An evaluation of BrailleTouch: mobile touchscreen text entry for the visually impaired. In: Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services, New York, NY, USA, pp. 317–326 (2012)
15. Baas, M., Guerrier, Y., Kolski, C., Poirier, F.: "Système de saisie de texte visant à réduire l'effort des utilisateurs à handicap moteur," In: Ergo'IA 2010: Proceedings of the Ergonomie et Informatique Avancee Conference, New York, NY, USA, pp. 19–26 (2010)
16. Wandmacher, T., Antoine, J.-Y., Poirier, F.: SIBYLLE: a system for alternative communication adapting to the context and its user. In: Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility, New York, NY, USA, pp. 203–210 (2007)
17. Coutaz, J., Calvary, G.: HCI and Software Engineering for User Interface Plasticity. In: Jacko, J.A. (ed.) Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, 3rd edn., pp. 1195–1220. CRC Press (2012)
18. Kubicki, S., Lepreux, S., Kolski, C.: RFID-driven situation awareness on TangiSense, a table interacting with tangible objects. Pers. Ubiquitous Comput. 16(8), 1079–1094 (2012)
19. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. 47(6), 381–391 (1954)
20. Lamy, J.-B., Venot, A., Bar-Hen, A., Ouvrard, P., Duclos, C.: Design of a graphical and interactive interface for facilitating access to drug contraindications, cautions for use, interactions and adverse effects. BMC Med. Inform. Decis. Mak. 8(1), 21 (2008)
21. Lamy, J.B., Duclos, C., Bar-Hen, A., Ouvrard, P., Venot, A.: An iconic language for the graphical representation of medical concepts. BMC Med. Inform. Decis. Mak. 8, 16 (2008)

Application of Frontal EEG Asymmetry to User Experience Research

Jing Chai[1,2], Yan Ge[1,*], Yanfang Liu[3], Wen Li[3], Lei Zhou[3], Lin Yao[3], and Xianghong Sun[1]

[1] Key Laboratory of Behavioral Science, Institute of Psychology, CAS
{chaij,gey,sunxh}@psych.ac.cn
[2] University of Chinese Academy of Sciences
[3] User Experience Lab, China Mobile Research Institute, Beijing, China
{liuyanfang,liwen,zhoulei,yaolin}@chinamobile.com

Abstract. The electrophysiology technique now provides an alternative way to evaluate users' emotional states in real time, but how to confirm the valence of emotions using these techniques is still a concern to researchers. Frontal alpha asymmetry (FAA) is often used as an index of pleasantness or liking in neuromarketing, but results in related fields are not consistent. In this study, we investigated the emotional states of users interacting with mobile phone applications (APPs) using FAA. Twenty participants participated in this experiment. They were asked to complete several tasks in a scene of everyday life using three APPs of the same type. EEG data and subjective evaluations were recorded during the experiment. The FAA results showed a positive trend when using an APP that provided an excellent user experience. The mechanism of emotional change during interacting with mobile applications and the implications of this research are also discussed in this study.

Keywords: user experience, emotional state, FAA, EEG.

1 Introduction

According to ISO 9241-210 [1], user experience (UX) is defined as "a person's perceptions and responses that result from the use or anticipated use of a product, system, or service." This concept represents the integrated feelings users get while using a product, including the way they understand a product (understanding experience) and their emotional response to it (emotional experience) along with the pleasure experienced from sensory perceptions (esthetic experience) [2]. The emotional experience is frequently measured by subjective evaluations in usability testing. The self-report method can be influenced by social desirability. This evaluation was often performed after participants completed the tasks, making it difficult to measure changes in emotional states in real time [3]. The electrophysiology technique provides an alternative solution for this problem. Skin conductivity, heart rate, blood pressure, etc. are

* Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Chaoyang Dist., Beijing, China, 100101.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 234–243, 2014.

effective indices for measuring arousal levels during human–machine interactions [4-6], but it is difficult to distinguish the valence of emotions during the interactions using these indices.

The electroencephalography (EEG) has been used in UX fields to measure emotional experiences and cognitive workloads, with α and β as the main indicators [7-12]. Li, et al. recorded EEG data when users were visiting a distance learning simulation website. The results showed a correlation between EEG-α rhythm (8-13 Hz) and emotional state. The amplitude, amplitude's deviation, and power of α wave decreased when users were reading simple and interesting content rather than boring content [8]. Stickel, et al. also used EEG recording in usability testing and analyzed several specific frequency bands including β (12.5–28 Hz) and α (8–12.5 Hz). This research revealed that learnability can be estimated by brainwave patterns: the α waves were dominant when the software was easy to learn; the β waves were dominant when the software was hard to learn [12]. EEG spectral analysis can also be used to assess the effects of pre-training on learning software. Masaki, et al. found that the β/α power decreased with the software use experience, in which they took β/α value as an index of mental workload. If the β/α value was greater than 1.0, it indicated a state of higher mental workload; otherwise, a value lower than 1.0 indicated a state of lower mental workload [9].

Although early studies found that the EEG could be a common indicator of cognitive workloads and emotional states, how to identify the valence of emotions when users interacted with a product remains an open question. The correlation between the hemispheric asymmetries in pre-frontal activity and approach withdraw related motivation is extensively used in basic research [13]. The frontal alpha asymmetry (FAA) is a potential and valuable index in this research field when used to index frontal brain activity when processing different states of emotion. In Davison's model [14], the left pre-frontal cortex (PFC) is involved in the processing of positive affects, whereas the right PFC is involved in the processing of negative affects. This index has been widely used in neuromarketing [10, 11]. Ohme, et al. found that FAA can capture the difference between two slightly different versions of TV ads in a few seconds [11]. It has also been used to explore preferences in product design [15-17].The FAA provides a new solution to the measurement problem of human–machine interaction.

The main purpose of this research is to investigate the feasibility of using neurophysiology index to evaluate emotional experiences in UX, with FAA as the main index of affects valence. Three different mobile applications were tested in the experiment, and all the behavior, subjective, and physiological data were collected. We assume a consistency between self-report users of experiences and frontal asymmetric activities, which will make a distinction between positive and negative emotion.

2 Methods

2.1 Participants

Twenty participants were recruited from an open access, online part-time job platform in China (http://zhan.renren.com/jobcome).They aged from 21 to 29 years old (mean: 23.8 ± 2.484), including ten females and ten males. Their experience of using this type of APP and emotional sensitivity were controlled.

2.2 Materials

Mobile applications. Three APPs, designated DZD, YEW, and HMB, were chosen as experimental materials according to their usability testing results in pilot studies. These APPs have similar functions, but different usability ratings. They provided a lot of information about restaurants, hotels, and entertainment venues in a city. Users could search, order, or evaluate a restaurant using APPs like these. The latest versions of all APPs were installed on a smart phone using an Android operating system in advance of the experiment.

Questionnaires. Three self-report questionnaires were used as subjective measurement in this study.

The Positive and Negative Affect Scale (PANAS). This scale was used to measure the current mood of participants. According to Watson and Tellegen's two-factor model, positive affect (PA) and negative affect (NA) are two basic and mutually independent dimensions in the structure of emotion [18]. PA refers to people's feelings of enthusiasm, activeness, and pleasure. A high PA value represents a state of powerfulness, concentration, and joviality. The opposite, NA relates to distress and unpleasantness, including anger, contempt, disgust, guiltiness, fear, and tension. A low NA value represents peace and calmness. The PANAS consists of 20 affective words, 10 for each dimension. The participants were to indicate to what extent their state matched each word on a scale from 1 ("very slightly or not at all") to 5 ("very much") [19].

User Experience Questionnaire (UEQ). UEQ is a widely used and most common tool for software quality and usability assessment. It can be used to assess the comprehensive impression of user experience in a convenient and quick way. It consists of six relatively independent factors, including attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty [20]. This scale includes 26 items, each composed of an adjective and its antonym. Users evaluated their preference between each pair of words using a 7-point scale.

System Usability Scale (SUS). SUS is used for usability assessment of software systems, products, or websites, especially for competitive analysis [21]. SUS includes 10 declarative sentences. The participants rated each sentence from 1 (not at all) to 5 (very much) according to how much they agreed with those sentences. Note that the SUS score only represents the overall usability of a system, and different attributes of the system such as effectiveness and efficiency are not measured in the SUS.

2.3 Procedures

When the participants arrived at the laboratory room, the moderator introduced them to the procedures and tasks of the experiment in detail and then prepared them for EEG recording. Next, the participants were asked to complete the PANAS for the first time, which would be used as the baseline of affective state for the whole experiment.

In the formal study, the participants used three APPs respectively to perform a series of tasks. The order of using APPs was counterbalanced in a Latin square sequence. For each APP, the order of tasks was consistent. First, the participant sat quietly and statically for three minutes. The EEG data of this stage was recorded as a baseline. Then, a daily situation was setup as follows, "Suppose a friend of yours has recently come to visit this city. Please book a restaurant and a hotel room for this friend according to the experimenter's instructions." The participant searched for related information and made choices using an APP. When the participant finished a task, she/he reported their choice orally and answered several questions asked by the experimenter, e.g., the reasons for making such choices and the difficulty of the task. If the participant felt the task was extremely difficult to complete after several minutes, she/he could give up. There were no strict time constraints for each task. When a task was executed for over ten minutes, the participant received a warning. The completion rate, completion time, and the difficulty of the task reported by the participants were collected as the behavioral data. After the tasks for one APP were finished, the participants were asked to complete two questionnaires, the PANAS and UEQ. The EEG data was collected for the entire session. The entire assessment for one APP lasted for about 30 min. After a brief break, participants continued on to the second APP. The situations and tasks for the three APPs were similar, except the meeting locations for each APP were different. When the tasks for all three APPs were completed, participants compared the usability of the APPs using the SUS. The entire experiment lasted approximately two hours.

2.4 EEG Recording and Analysis

The EEG data was continuously recorded from 32 scalp sites using tin electrodes mounted in an elastic cap arranged according to the 10–20 international placement system (Neuroscan Inc.). The online reference was right mastoid (A2). The electrode sites on the participant's mastoids and forehead were cleaned with alcohol cotton balls gently. The impedances of the EEG electrodes were below 5 kΩ. The EEG data were amplified with a bandpass filter of 0.05–100 Hz and digitized at 500 Hz. The recordings obtained from the prefrontal and frontal regions of the cortex (Fp1, Fp2, F3, F4, F7, and F8) were saved.

The EEG data was processed using Neuroscan 4.5 software. All the EEG data were DC corrected and re-referenced to linked mastoids offline. The filter was set to 30 Hz low pass and 0.1 Hz high pass. Then, the EEG data were epoched into periods of 512 points (i.e., 1024 ms). The power of alpha band (8–12 Hz) and beta band (12–29.3 Hz) in each of the recorded electrodes was calculated for further analysis.

The frontal alpha asymmetry (FAA) index was calculated as the difference between right-hemispheric data minus left-hemispheric data (ln (right alpha power) - ln (left alpha power)) according to previous studies [22, 23]. Due to the negative correlation between alpha power and brain activation, the positive score of the FAA index implies the dominance of left PFC and the negative score of the FAA index implies the dominance of right PFC.

3 Results

All the data were processed by SPSS 16.0. The EEG data exceeding three standard deviations were considered as extreme values and removed from further analyses.

3.1 Behavioral Results

The behavioral data included the task completion rate, the task completion time, and the task difficulty, which were recorded by the experiment assistant during the experiment. A single-factor repeated measure of variance analysis was taken to analyze the difference among the three APPs. There are significant differences among the three APPs in the completion rate ($F = 44.333$, $p < 0.01$), completion time ($F = 12.314$, $p < 0.01$), and task difficulty ($F = 112.405$, $p < 0.01$). Further analysis showed that the completion rate of HMB was significantly lower than that of YEW ($p < 0.01$) and DZD ($p < 0.01$); the completion time of HMB was significantly longer than that of YEW ($p < 0.01$) and DZD ($p < 0.01$), and the self-report difficulty of HMB was significantly higher than that of YEW ($p < 0.01$) and DZD ($p < 0.01$). No significant difference was found between YEW and DZD in completion time ($p = 0.152$) or completion rate ($p = 0.163$). But YEW is significantly more difficult than DZD ($p < 0.01$). The descriptive statistics of the behavioral data are given in Table 1.

Table 1. Descriptive statistics of the behavioral data

	YEW (M ± SD)	DZD (M ± SD)	HMB (M ± SD)	*F*
Completion rate	0.950 ± 0.215	1.000 ± 0.000	0.500 ± 0.498	44.333**
Completion time (100s)	1.722 ± 1.038	1.437 ± 1.019	2.764 ± 1.612	12.314**
Task difficulty	2.025 ± 0.898	1.575 ± 0.794	3.775 ± 1.078	112.405**

* $p < 0.05$, ** $p < 0.01$

3.2 Subjective Evaluation

UEQ and SUS. Results from repeated measures of variance analysis showed a significant main effect for the three APPs ($F = 94.162$, $p < 0.01$) in UEQ. Results also revealed a significant main effect for the three APPs ($F = 113.274$, $p < 0.01$) in SUS. Pairwise comparisons showed that the SUS score of HMB was lower than that of YEW ($p < 0.01$) or DZD ($p < 0.01$). No significance was found between YEW and DZD in the SUS score ($p = 0.07$).

Table 2. Descriptive statistics of UEQ and SUS

	YEW (M ± SD)	DZD (M ± SD)	HMB (M ± SD)	*F*
UEQ	1.141 ± 1.064	1.465 ± 0.737	−1.072 ± 0.946	94.162**
SUS	82.375 ± 16.130	76.375 ± 15.780	20.125 ± 11.711	113.274**

* p < 0.05, ** p < 0.01

PANAS. The positive affect (PA) and the negative affect (NA) were analyzed respectively by repeated measures of variance analysis. Results showed significant differences among the three APPs in both PA ($F = 44.457$, $p < 0.01$) and NA ($F = 17.341$, $p < 0.01$). Pairwise comparisons showed that the PA value when using HMB was lower than when using YEW ($p < 0.01$) or DZD ($p < 0.01$), the NA value when using HMB was significantly higher than when using YEW ($p < 0.01$) or DZD ($p < 0.01$). No significance was found between YEW and DZD neither in PA value ($p = 0.702$) nor NA value ($p = 0.07$). The descriptive statistics of PANAS are given in Figure 1.

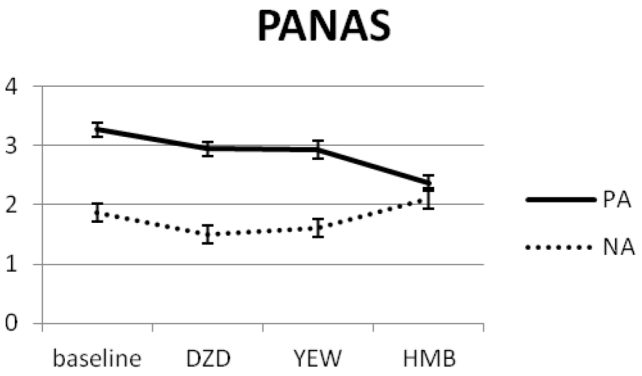


Fig. 1. Descriptive statistics of PANAS. Standard error bars are included

3.3 FAA Analyses

We only analyzed the data of the first two tasks for each APP. The results of FAA are shown in Figure 2. Unfortunately, results from repeated measures of variance analysis showed no significance in FAA index among the three APPs ($F = 1.417$, $p = 0.261$). But we could see a positive trend in DZD in comparison to the other two APPS.

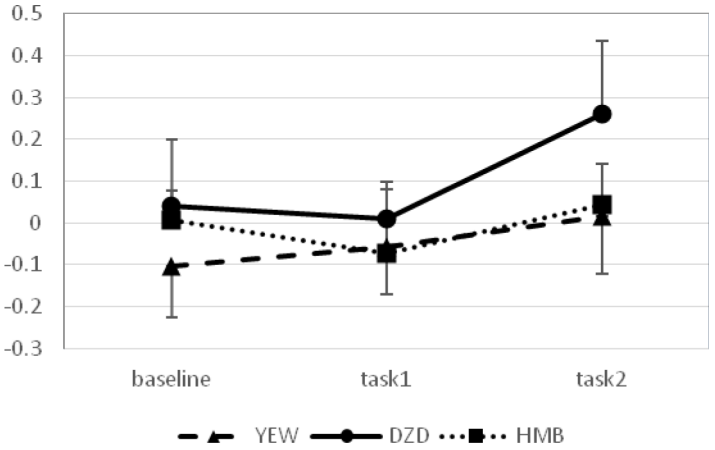


Fig. 2. Descriptive statistics of FAA index. Standard error bars are included.

3.4 Correlations

The correlations between EEG data and subjective data were also analyzed. Results of Pearson correlation coefficients revealed that the PA value is negatively correlated with both alpha power ($r = -0.310$, $p < 0.05$) and alpha power's change from baseline (i.e., task–baseline, $r = -0.309$, $p = 0.026$), and the NA change from baseline (task–baseline) is negatively correlated with both beta power ($r = -0.298$, $p < 0.05$) and beta power's change from baseline ($r = -0.318$, $p < 0.05$). The correlation between the FAA change (task–baseline) and the task's difficulty are marginally significant ($r = -0.264$, $p = 0.053$). Also in task 1, the results revealed an inverse correlation between the FAA index and the task completion time ($r = -0.315$, $p < 0.05$).

4 Discussion and Conclusions

This experiment investigated the feasibility of using EEG to distinguish positive and negative emotional states in user experiences. The user experience was significantly different for these three APPS. First, the results of the behavioral data showed YEW and DZD were better than HMB in task completion rate, completion time, and self-report task difficulty. Second, YEW and DZD were better than HMB in usability; their scores on UEQ and SUS were higher than the scores for HMB. Finally, the emotional state after using each APP was also different. The PA after using HMB was lower than for the other two APPS, and the trend of NA was reversed. These results revealed that participants felt more positive emotions and less negative emotions when using DZD and YEW than when using HMB.

No significant difference was found in FAA among the three APPs, but there were some trends that we can discuss. The FAA was positive when using DZD, revealing an approach trend with this APP. This result was consistent with behavioral and

subjective evaluation results. The FAAs were around zero when using YEW and HMB and revealed no preference in using these two APPs, even the results of behavior and subjective evaluation showed YEW was better than HMB. It is worth noting that the task difficulty of YEW was greater than for DZD based on the self-report results. The participants felt different moods when using these two APPs, but they could not consciously report this difference. So, one possible explanation is that the EEG results were probably to distinguish the tiny differences in emotional states. The marginally significant correlation between FAA change and task difficulty supported this explanation.

The results from the correlation analysis also indicated a significant relationship between EEG data and subjective evaluation. Previous studies have proven that brainwave activity relates to changes in mental or physical states, that is, the dominance of fast rhythmic activity (beta/gamma) indicates states of high arousal (e.g., reasoning, problem solving) and the dominance of slow rhythmic activity (alpha) indicates a relaxed state [12]. On one hand, a high PA value represents a state of powerfulness, concentration, and joviality [19]. The PA score was negatively correlated with alpha power and the change of alpha power in this experiment. It implies that less concentration accompanies high alpha power. On the other hand, low NA represents a state of peace and calmness [19]. A negative correlation between the change of NA and beta power was found, but the internal mechanism still needs more discussion.

There are some limitations in this research. The emotional arousal levels for the three apps are not high enough. This could be a reason that no significant difference was found in FAA among three APPs. The materials used in marketing and design were complicated, which often arouses people's emotional response in physical or visual ways [10, 11, 16]. However, the interfaces of the three APPs used here were much simple than those in commercial ads. The tasks, such as finding a restaurant or planning a route to a specific location, were too easy to cause a change in emotions. The differences in usability were easy for users to perceive and self-report, but the change of affect caused by the usability levels of the APPs was extremely slight to be captured by the FAA. Other software, which could arouse strong fluctuations, such as game APPs, can be explored in future research.

In sum, using EEG indicators to evaluate the emotional state of users is a feasible tool in human computer interaction. FAA could be used as an index of approach, although we did not reach significant results in this study. The stability of FAA and its scope of application need to be explored in future research.

Acknowledgement. This work was supported by User Experience Lab of China Mobile Research Institute, Science and Technology (S&T) basic work (2009FY110100) and NSF China (31100750, 91124003).

Reference

1. FDIS, I.: 9241-210: 2009 Ergonomics of human system interaction-Part 210: Human-centered design for interactive systems (2009)
2. Hekkert, P.: Design aesthetics: principles of pleasure in design. Psychology Science 48, 157 (2006)
3. Vermeeren, A.P., Law, E.L.-C., Roto, V., Obrist, M., Hoonhout, J., Väänänen-Vainio-Mattila, K.: User experience evaluation methods: current state and development needs. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, pp. 521–530. ACM (2010)
4. Riseberg, J., Klein, J., Fernandez, R., Picard, R.W.: Frustrating the user on purpose: using biosignals in a pilot study to detect the user's emotional state. In: CHI 1998 Conference Summary on Human Factors in Computing Systems, pp. 227–228. ACM (1998)
5. Scheirer, J., Fernandez, R., Klein, J., Picard, R.W.: Frustrating the user on purpose: a step toward building an affective computer. Interacting with Computers 14, 93–118 (2002)
6. Wilson, G.M., Sasse, M.A.: From doing to being: getting closer to the user experience. Interacting with Computers 16, 697–705 (2004)
7. Hu, J., Nakanishi, M., Matsumoto, K.-I., Tagaito, H., Inoue, K., Shima, K., Torii, K.: A method of usability testing by measuring brain waves. IEEE Trans. Software Eng. 25, 474–491 (1999)
8. Li, X., Hu, B., Zhu, T., Yan, J., Zheng, F.: Towards affective learning with an EEG feedback approach. In: Proceedings of the First ACM International Workshop on Multimedia Technologies for Distance Learning, pp. 33–38. ACM (2009)
9. Masaki, H., Ohira, M., Uwano, H., Matsumoto, K.-I.: A Quantitative Evaluation on the Software Use Experience with Electroencephalogram. In: Marcus, A. (ed.) HCII 2011 and DUXU 2011, Part II. LNCS, vol. 6770, pp. 469–477. Springer, Heidelberg (2011)
10. Ohme, R., Matukin, M., Szczurko, T.: Neurophysiology uncovers secrets of TV commercials. Der Markt 49, 133–142 (2010)
11. Ohme, R., Reykowska, D., Wiener, D., Choromanska, A.: Analysis of neurophysiological reactions to advertising stimuli by means of EEG and galvanic skin response measures. Journal of Neuroscience, Psychology, and Economics 2, 21 (2009)
12. Stickel, C., Fink, J., Holzinger, A.: Enhancing Universal Access – EEG Based Learnability Assessment. In: Stephanidis, C. (ed.) HCI 2007. LNCS, vol. 4556, pp. 813–822. Springer, Heidelberg (2007)
13. Briesemeister, B.B., Tamm, S., Heine, A., Jacobs, A.M.: Approach the Good, Withdraw from the Bad—A Review on Frontal Alpha Asymmetry Measures in Applied Psychological Research. Psychology 4, 261–267 (2013)
14. Davidson, R., Schwartz, G., Saron, C., Bennett, J., Goleman, D.: Frontal versus parietal EEG asymmetry during positive and negative affect. Psychophysiology 16, 202–203 (1979)
15. Kline, J.P., Blackhart, G.C., Woodward, K.M., Williams, S.R., Schwartz, G.E.: Anterior electroencephalographic asymmetry changes in elderly women in response to a pleasant and an unpleasant odor. Biological Psychology 52, 241–250 (2000)
16. Park, M.-K., Watanuki, S.: Electroencephalographic responses and subjective evaluation on unpleasantness induced by sanitary napkins. Journal of Physiological Anthropology and Applied Human Science 24, 67–71 (2005)

17. Tomico, O., Mizutani, N., Levy, P., Takahiro, Y., Cho, Y., Yamanaka, T.: Kansei physiological measurements and constructivist psychological explorations for approaching user subjective experience during and after product usage. In: Proceedings of the DESIGN 2008, 10th International Design Conference, pp. 529–536 (2008)
18. Watson, D., Tellegen, A.: Toward a consensual structure of mood. Psychological Bulletin 98, 219 (1985)
19. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology 54, 1063 (1988)
20. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (ed.) USAB 2008. LNCS, vol. 5298, pp. 63–76. Springer, Heidelberg (2008)
21. Brooke, J.: SUS-A quick and dirty usability scale. Usability Evaluation in Industry 189, 194 (1996)
22. Harmon-Jones, E., Allen, J.J.: Behavioral activation sensitivity and resting frontal EEG asymmetry: covariation of putative indicators related to risk for mood disorders. Journal of Abnormal Psychology 106, 159 (1997)
23. Wheeler, R.E., Davidson, R.J., Tomarken, A.J.: Frontal brain asymmetry and emotional reactivity: A biological substrate of affective style. Psychophysiology 30, 82–89 (1993)

Theoretical Investigation on Disuse Atrophy Resulting from Computer Support for Cognitive Tasks

Kazuhisa Miwa and Hitoshi Terai

Graduate School of Information Science, Nagoya University
Nagoya, 464-8601, Japan
{miwa,terai}@is.nagoya-u.ac.jp

Abstract. We propose a new concept, disuse atrophy in cognitive abilities, i.e., cognitive disuse atrophy. Generally, the term "disuse atrophy" has been used to describe physical atrophy, such as muscle wasting. We advance the idea that disuse atrophy appears not only as physical loss but also as a loss of cognitive abilities. To understand the mechanisms underlying cognitive disuse atrophy, we note the duality of cognitive activities such as performance- and learning-oriented activities when engaging in tasks. It is crucial to investigate the balancing of these two types of activities as the assistance dilemma in learning science. We explored principles for controlling this balance based on two theories: cognitive load theory and goal achievement theory. Cognitive load theory distinguishes three types of cognitive loads. This theory proposes to suppress the extraneous load to the minimum, while assigning adequate amounts of the germane load for learning-oriented activities into working memory, and still leave enough resources for the intrinsic load of performance-oriented activities. Goal achievement theory assumes principles from the viewpoint of goal setting. Specifically, orientation to a performance goal activates performance-oriented activities, and orientation to a learning goal causes learners to direct their efforts to learning-oriented activities.

Keywords: Disuse atrophy, Assistance dilemma, Cognitive load theory, Goal achievement theory.

1 Introduction

1.1 Automated Systems as a Third Generation of Tools

Humans have acquired overwhelming abilities by developing a variety of tools for cultivating and extending their controlled world. Knives and hammers are representative examples of tools that have been used since ancient times. These tools are used to support physical human activities such as cutting, hitting, and building and are regarded as the first generation of tools in the history of tool development.

By the end of the twentieth century, new type of tools, known as cognitive artifacts have emerged in our society [19]. Cognitive artifacts are regarded as

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 244–254, 2014.

the second generation of tools. These tools support human cognitive activities and are often called "systems" because of their functions. These systems are significantly different from traditional tools in their functionality, which is referred to as multiple, high-level editing, and interactive functionality. This difference cause discordance between users' intentions and actual processing conducted in a system [20][24].

Yet another new type of system has recently emerged, i.e., automated systems that perform autonomous activities such as automatic driving vehicles and automatic cleaning robots. These tools have unique natures that are different from traditional tools that support human physical and cognitive activities [22][25]. These automated systems can be seen as the third generation of tools.

In the use of traditional tools and systems, humans have the primary role of performing a task, and the system is secondary in support of the activity. Even when large roles are assigned to systems in performing a task, the relationship between systems and humans has been characterized as a collaboration in which mutual interaction is a key factor for determining total performance. But in the use of automated systems, the system becomes the main actor and the task of the human is to monitor the behavior of the system.

Information processing by automated systems that carry out tasks rather than humans is extremely complex. Hence, the inner processing of the system is usually packaged as a black box, which is impossible to be understood by the user. In second generation systems, there is a similar but less developed aspect, so users must construct mental models to understand and interact with these systems. However, in third generation systems, users often entrust the entire operation to the system without even minimal understanding of the workings of the system.

1.2 Cognitive Disuse Atrophy

Automated systems undertake tasks in a variety of fields and greatly enhance human abilities for working. However, the convenience emerging from the use of such systems has, in some cases, caused negative impacts on human society. A majority of people may be experiencing these issues in daily life, e.g., difficulty in memorizing maps due to daily usage of a car navigation system or difficulty remembering the accurate spelling of words because of using a word processor with spell checker software. Initially, Norman highlighted issues related to the use of automated systems. Human factor studies have reported that the continuous use of automated systems decreases users' manipulation abilities [25] and, more seriously, complacency on this front causes aircraft accidents [31].

This study proposes a new concept, cognitive ability disuse atrophy, i.e., a loss of cognitive ability due to a lack of use. We see this as a key issue underlying some human factor problems that emerge when people engage in cognitive tasks in collaboration with computers. This study is a theoretical investigation of this issue. The term "disuse atrophy" is generally used for physical body atrophy, such as muscle wasting. When muscles are no longer in use, they slowly weaken. This weakening, or atrophy, can also occur from continuous physical support that

leads to a minimal use of the body. We advance the idea that disuse atrophy occurs not only in the physical body but also in cognitive ability.

1.3 Duality of Cognitive Activities

To understand the mechanisms underlying cognitive disuse atrophy, we begin with the duality of cognitive processing when engaging in a task. Generally, there are two objectives for performing a task. One ordinary objective is to perform and complete the task. However, there is another important objective, i.e., for performers to develop proficiency and knowledge by performing the task. Performance and mastery are the prime reasons to engage in a task. We contend that disuse atrophy emerges when the aspect of mastery is lost.

For example, consider car navigation systems. When a person searches for a route from a current location to a new destination, they usually try to remember a mental map, a configuration of the possible pathways, select candidate pathways related to the target route, and decide on the best route from multiple candidates while considering current traffic and construction. These cognitive information processing efforts develop mastery for remembering maps and acquiring the skills to search for a route. However, when we use a navigation system, we do not need to perform any such mental activities, because all of this type of processing is performed by the system. All the person has to do is enter the destination and press the confirmation button. From the viewpoint of performance that is all it takes to achieve the goal. But for remembering a map and becoming able to find a route with a printed map, such mental activities that lead to mastery are important. Because car navigation systems deprive users of making such efforts for mastery, they cause mental disuse atrophy.

2 The Assistance Dilemma

2.1 Performance- and Learning-Oriented Activities

Next, we investigated issues related to performance and mastery from the viewpoint of the duality of cognitive activities in learning. When students engage in learning through practice, they usually solve problems for exercise. Note that cognitive activities for solving problems are not necessarily equivalent to cognitive activities for learning. Consider a learning situation in which students learn procedural knowledge to solve mathematical problems. Students solve example problems by applying the procedural knowledge. To optimize the learning effects, students should examine the conditions under which the knowledge applies, relate the newly acquired knowledge to previously acquired knowledge while considering the relationship between the two knowledge sets, and finally, generalize the knowledge for application to a variety of problems. This paper defines the former cognitive activities related to problem solving as "performance-oriented activities" and the latter cognitive activities for learning as "learning oriented activities." Performance- and learning-oriented activities correspond to cognitive activity for performance and mastery as discussed previously. Both activities are required to maximize the effects of learning.

2.2 The Assistance Dilemma

The importance of balancing these two types of cognitive activities has been recognized as the assistance dilemma. Recent intelligent tutoring systems have highly interactive features. Such systems give participants a variety of feedback, such as verification, correct response, try again encouragement, error flagging, and elaboration messages. In this context, the assistance dilemma has been recognized. Koedinger and Aleven (2007) pointed out a crucial question: How should learning environments balance assistance giving and withholding to achieve optimal learning? [13] The problem is that high assistance sometimes provides successful scaffolding and improves learning; but at other times, it elicits superficial responses given without consideration. On the other hand, low assistance sometimes encourages students to make great efforts in learning, but other times it results in enormous errors and interferes with effective learning. To solve this dilemma, the levels of support in tutoring systems must be adaptively controlled.

Figure 1 illustrates the assistance dilemma. Consider a student who is solving problems for exercise while receiving assistance from a tutoring system. The horizontal axis represents the level of support, defined by the amount of help information. The level of support is at the minimum when the student solves the problems by themselves without system assistance. The level of support is at the maximum when the student is given a final solution directly. The latter is called a bottom out hint. The vertical axis shows the problem solving performance during the learning phase and the achievement level of learning that is measured after the learning phase. The former is usually evaluated by the solution time, the rate of incorrect answers, and the errors committed in solving problems while assistance is being given. The latter is evaluated by post-test scores when problems are solved by the students themselves without assistance.

Figure 1 shows that as the level of support increases, problem solving performance also increases constantly; but the post-test scores indicate that the learning effect reaches its peak at the optimum assistance level and begins to decrease from that point, meaning that over-assistance occurs in the left side of the graph.

The assistance dilemma occurs in the left side of the graph. The dilemma is caused by the duality of cognitive processing. Tutoring support activates performance-oriented activities and improves problem solving performance, but over support contradictorily inhibits learning- oriented activities, thus causing the assistance dilemma. Now we go back to the issue of disuse atrophy. Automated systems perform everything. Humans do not need to do anything. From the viewpoint of the assistance dilemma framework shown in Figure 1, we infer that support level almost reaches the maximum at the leftmost side of the graph resulting in the emergence of cognitive disuse atrophy.

2.3 Empirical Evidence of Dilemma

Miwa, et al. empirically confirmed the dilemma of the performance and learning curves depicted in Figure 1 using two experimental tasks [16]. One was the Tower

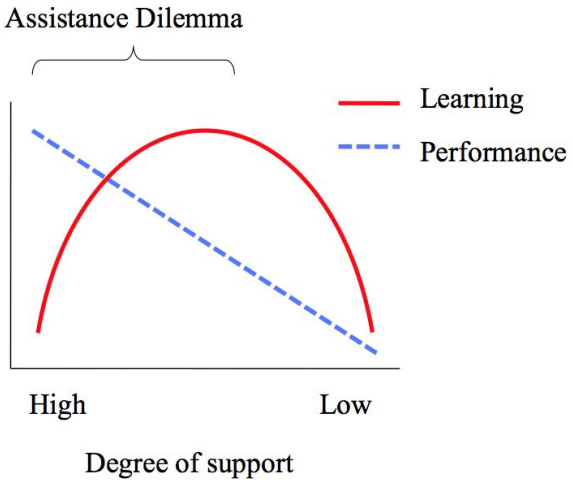


Fig. 1. The Assistance dilemma and performance and learning curves as a function of the levels of support

of Hanoi (TOH) puzzle, which is a representative experimental task widely used in problem-solving studies. The other was a natural deduction (ND) task. TOH is a simple task in which the problem space is systematically organized and is not very large. Problem solving is achieved by only one operator that corresponds to disk movement. The knowledge and strategies for the solution are represented by less than ten production rules, whereas ND is a more complex task in which problem space is much larger than that of TOH. To solve problems, since participants must acquire many kinds of inference rules and solution strategies, a complete model for solving ND problems consists of around a hundred production rules. The assistance dilemma was confirmed in both such relatively different types of tasks.

3 Cognitive Load Theory

The balance of performance- and learning-oriented activities is critical to solving the problem of cognitive disuse atrophy. How can we control the balance of the two activities? To answer this question, we examine two theories: cognitive load theory constructed in cognitive and instructional sciences and goal achievement theory developed mainly in educational psychology. Cognitive load theory (CLT) has provided informative perspectives for designing learning environments based on the constraints of cognitive architecture. In this study, we reinterpret the relative findings of CLT for balancing learning- and performance-oriented activities.

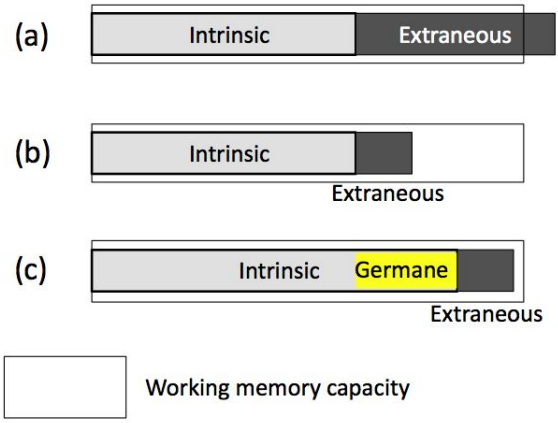


Fig. 2. Relation of intrinsic, extraneous, and germane loads

3.1 Three Types of Cognitive Loads

CLT distinguishes three types of cognitive loads: intrinsic, extraneous, and germane [27][29]. The intrinsic load is defined as the basic cognitive load required to perform a task. As the difficulty of the task increases and the degree of expertise of the performer decreases, the intrinsic load increases. The extraneous load is defined as wasted cognitive load that does not relate to learning activities, but emerges reluctantly. One reason that the extraneous load occurs is due to the inappropriateness of learning material designs. For example, when related information is not arranged properly, the extraneous load is increased by the efforts of doing irrelevant searches to gather the related information. The germane load is defined as the load used for learning, such as for constructing schemata activities.

Figure 2 illustrates the relationship among the three cognitive loads [29]. Figure 2 (a) illustrates the status in which the cognitive load exceeds the limits of the performer's working memory capacity due to the increase in the extraneous load. In this overloaded situation, learners make enormous errors, spend too much time performing the task, and occasionally, may not be able to perform the task. Figure 2 (b) shows cognitive loads that fall within a range where learners perform a task easily and show good results. CLT proposes that in such situations with memory capacity to spare, it is important to raise the germane load to activate learning activities, as illustrated in Figure 2 (c).

3.2 Controlling Cognitive Loads

It is assumed that the intrinsic load occurs to perform performance-oriented activities and the germane load increases to perform learning-oriented activities. CLT has revealed how these loads vary with changes in the degree of learning support.

Adequate support lets learners perform a task easily and suppresses the intrinsic load. When materials are properly designed, the emergence of the extraneous load is also minimized and the working memory has plenty of room to perform the task. This is advantageous to performance-oriented activities, as illustrated in Figure 2 (b). To activate learning-oriented activities, the germane load should be increased. To do so, learners are properly guided to engage in learning-oriented activities, increasing the germane load while the extraneous load is still minimized, as illustrated in Figure 2 (c).

3.3 List of Design Principles

A list of principles for designing learning materials and systems to minimize the extraneous load and increase the germane load has been developed by CLT [27].

- **The Goal-free effect**: Set up a situation in which learners search for various solutions. To let learners to do so, do not provide a specific goal that drives learners into searching for a single specific solution.
- **The Worked example effect**: Worked examples are used as an alternative to learning by problem solving [26][21]. Worked examples are defined as a solution example that includes steps to solve the problem, equations for solving the problem, and the final solution. Learners are guided to follow the process of experts' problem solving by tracing their steps in the worked examples. Learning by worked examples minimizes the learners' extraneous load and provides memory space for assigning the germane load.
- **Spirit attention effect**: When relevant information is presented in a fragmented way, the extraneous load increases in the search for mutual references. Suppress the extraneous load by arranging related information together, e.g., including explanatory text in a figure.
- **The Modality effect**: In information presentation, combine multiple modalities to suppress a split-attention effect. For example, use auditory guidance for presentation with text information.
- **Redundancy effect**: Irrelevant information is separately presented no to be noted by learners because if such information is referred, the extraneous load increases.
- **Variability effect**: Let learners investigate a single topic with multiple representations and under various contexts, activating their cognitive efforts for generalizing knowledge and encouraging schema creation.

4 Goal Achievement Theory

4.1 Performance Goal and Learning Goal

Another decisive factor that determines whether learners assign cognitive efforts to performance- or learning-oriented activities are the goals they set for engaging in the task. Goal achievement theory (GAT) has provided theoretical perspectives on the relationship between the goals that students set for themselves and

their learning activities. It has also accumulated a vast amount of empirical findings.

In GAT, student goals are divided into learning goals and performance goals [4]. Learning goals motivate students to aim for developing their own abilities, but performance goals are motivated by a desire to seek higher social evaluation, rather than their own development. This implies that the former goal activates learning-oriented activities and the latter activates performance-oriented activities. Similar goals that may relate to both goals have been investigated from a variety of viewpoints in various contexts. Table 1 shows a summary of a set of families of performance and learning goals [30][17][18][12][15][2][14]. Extensive meta-analysis was conducted about the relationships of such goals, each of which is called by a different name, but some of which share common properties[11].

Table 1. Distribution of the numbers of participants who drew the reversed figure in Experiment 1

Learning	Performance	
Intrinsic	Extrinsic	Lepper, Corpus, & Iyengar, 2005 Vansteekiste, Lens, & Deci, 2006
Task-involved	Ego-involved	Nicholls, 1984 Nolen, 1988 Jagacinski & Nicholls, 1987
Task	Ability	Midgley, et al., 1998
Mastery	Performance	Ames & Archer, 1988

In the early stages of the study, GAT findings stressed the superiority of learning goals over performance goals. Specifically, Utman (1997) confirmed, through meta-analysis of preceding studies, that the priority of the learning goal is more significant when performers are adult and tasks are complex [28]. Through the history of investigation, the shift from performance goals to learning goals has been regarded as adaptive development through which learners become challenged and exhibit higher independence in performing a task [1]. Meanwhile, it has been confirmed that when learners set performance goals that do not challenge them to new missions, they give up in the face of difficult requirements even though they are relatively proactive when they receive high evaluation.

Insert Table 1 about here Table 1: Family of the learning and performance goals.

4.2 2 x 2 Framework

After 2000, an additional dimension, defined as approach and avoidance status, began to be considered in the distinction of learning and performance goals, proposing a 2 x 2 framework [5][23][10]. Approach status is the orientation toward reaching high scores. Avoidance status is the orientation toward evading getting low scores.

In this framework, the disadvantage of the performance goal becomes prominent in the avoidance status. Rather, the performance goal in the approach status is regarded as a desirable attitude for achieving high learning effects. In the past, it has usually been confirmed that performance goals motivate more superficial cognitive processing, while learning goals activate deeper processing. However, Elliot, et al. indicated that in the approach status, orientation to the performance goal is positively correlated with test scores, even though in the avoidance status, it is negatively correlated [6]. Harackiewicz, et al. indicated that based on their meta-analysis of empirical studies, performance goals in the approach status lead to positive impacts on the improvement of task performance [9].

4.3 Controlling Goals

Although GAT did not intend to provide methods on how to control goal setting, some of the studies consider this point. One set of studies indicated the possibility that, with instruction, students' goal orientation can be controlled [8][7]. These studies showed that when the point where knowledge and skills are developed through achieving a task is stressed, learners are motivated to set learning goals. When one student's performance is compared to another, students are motivated to set performance goals. Another set of studies, e.g., Ames (1992) and Church, et al. (2001), explored goal setting from the viewpoint of curriculum design in classroom settings in which practical principles were used to guide students to orient toward setting learning goals [1][3]. Research indicated the possibility that student goal setting may be manipulated by several factors, such as the nature of study requirements, how they were being evaluated, how responsibility was assigned, and voluntary attitudes of students.

In GAT, performance and learning goals are regarded as decisive factors for assigning learners' cognitive efforts to performance- and learning-oriented activities. The accordance and conflict aspects for both goals reveal central issues of balancing learning- and performance- oriented activities.

5 Toward a Solution for Cognitive Disuse Atrophy

This paper confirmed that the duality of cognitive activities underlies the emergence of cognitive disuse atrophy. Specifically, atrophy is caused when performance-oriented activities are assigned for performing a task without the learning-oriented activities. This insight reveals that the balance of the two types of cognitive activities is crucial to avoid cognitive disuse atrophy. Two theories, CLT and GAT, challenge this issue from different stand points. CLT is usually used to assess issues of design principles for constructing learning materials and environments. Consequently, the investigation of the relationship between the two cognitive activities is explored from the outside environment. To maximize learning effects by manipulating the two cognitive activities, the extraneous load must be suppressed to the minimum and an adequate amount of the germane load must be assigned for

learning-oriented activities into working memory, while reserving enough resources for the intrinsic load needed for performance-oriented activities. There are various principles proposed for doing so. However, GAT originates partially from personality psychology. Therefore, GAT theory has provided findings for controlling the two types of cognitive activities from inside the performers. GAT establishes perspectives for understanding the two types of cognitive activities from the goal setting perspective. Specifically, orientation to performance goals activates performance-oriented activities, while orientation to learning goals motivates learners to direct their efforts to learning-oriented activities. This paper proposes a first step for solving cognitive disuse atrophy based on the two primary cognitive theories that have been established from both theoretical and empirical perspectives.

Acknowledgement. This research was partially supported by HAYAO NAKAYAMA Foundation for Science & Technology and Culture.

References

1. Ames, C.: Classrooms: Goals, structures, and student motivation. Journal of Educational Psychology 84(3), 261–271 (1992)
2. Ames, C., Archer, J.: Achievement goals in the classroom: Students' learning strategies and motivation processes. Journal of Educational Psychology 80(3), 260–267 (1988)
3. Church, M.A., Elliot, A.J., Gable, S.L.: Perceptions of classroom environment, achievement goals, and achievement outcomes. Journal of Educational Psychology 93(1), 43–54 (2001)
4. Dweck, C.S.: Motivational processes affecting learning. American Psychologist 41(10), 1040–1408 (1986)
5. Elliot, A.J., Church, M.A.: A hierarchical model of approach and avoidance achievement motivation. Journal of Personality and Social Psychology 72(1), 218–232 (1997)
6. Elliot, A.J., McGregor, H.A., Gable, S.: Achievement goals, study strategies, and exam performance: A mediational analysis. Journal of Educational Psychology 91(3), 549–563 (1999)
7. Elliott, E.S., Dweck, C.S.: Goals: an approach to motivation and achievement. Journal of Personality and Social Psychology 54(1), 5–12 (1988)
8. Graham, S., Golan, S.: Motivational influences on cognition: Task involvement, ego involvement, and depth of information processing. Journal of Educational Psychology 83(2), 187–194 (1991)
9. Harackiewicz, J.M., Barron, K.E., Pintrich, P.R., Elliot, A.J., Thrash, T.M.: Revision of achievement goal theory: Necessary and illuminating. Journal of Educational Psychology 94, 638–645 (2002)
10. Harackiewicz, J.M., Linnenbrink, E.A.: Multiple achievement goals and multiple pathways for learning: The agenda and impact of paul r. pintrich. Educational Psychologist 40(2), 75–84 (2005)
11. Hulleman, C.S., Schrager, S.M., Bodmann, S.M., Harackiewicz, J.M.: A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? Psychological Bulletin 136(3), 422–449 (2010)

12. Jagacinski, C.M., Nicholls, J.G.: Competence and affect in task involvement and ego involvement: The impact of social comparison information. Journal of Educational Psychology 79(2), 107–114 (1987)
13. Koedinger, K.R., Aleven, V.: Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review 19, 239–264 (2007)
14. Lepper, M.R., Corpus, J.H., Iyengar, S.S.: Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. Journal of Educational Psychology 97(2), 184–196 (2005)
15. Midgley, C., Kaplan, A., Middleton, M., Maehr, M.L., Urdan, T., Anderman, L.H., Anderman, E., Roeser, R.: The development and validation of scales assessing students' achievement goal orientations. Contemporary Educational Psychology 23(2), 113–131 (1998)
16. Miwa, K., Terai, H., Nakaike, R.: Tradeoff between problem-solving and learning goals: Two experiments for demonstrating assistance dilemma. In: Proceedings of 34rd Annual Conference of the Cognitive Science Society, pp. 2008–2013 (2012)
17. Nicholls, J.G.: Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. Psychological Review 91(3), 328–346 (1984)
18. Nolen, S.B.: Reasons for studying: Motivational orientations and study strategies. Cognition and Instruction 5(4), 269–287 (1988)
19. Norman, D.A.: Cognitive artifacts. In: Carroll, J.M. (ed.) Designing interaction: Psychology at the Human-Computer Interface. Cambridge University Press (1991)
20. Norman, D.A.: The design of everyday things. Basic Books (2002)
21. Paas, F.G., Van Merriënboer, J.J.: Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. Journal of Educational Psychology 86(1), 122–133 (1994)
22. Parasuraman, R., Riley, V.: Humans and automation: Use, misuse, disuse, abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society 39(2), 230–253 (1997)
23. Pintrich, P.R.: An achievement goal theory perspective on issues in motivation terminology, theory, and research. Contemporary Educational Psychology 25(1), 92–104 (2000)
24. Rasmussen, J.: Information Processing and Human-Machine Interaction. An Approach to Cognitive Engineering. North-Holland (1986)
25. Sarter, N.B., Woods, D.D.: Team play with a powerful and independent agent: a full-mission simulation study. Human Factors: The Journal of the Human Factors and Ergonomics Society 42(3), 390–402 (2000)
26. Sweller, J., Cooper, G.A.: The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction 2(1), 59–89 (1985)
27. Sweller, J., Van Merrienboer, J.J., Paas, F.G.: Cognitive architecture and instructional design. Educational Psychology Review 10(3), 251–296 (1998)
28. Utman, C.H.: Performance effects of motivational state: A meta-analysis. Personality and Social Psychology Review 1(2), 170–182 (1997)
29. Van Merriënboer, J.J., Sweller, J.: Cognitive load theory in health professional education: design principles and strategies. Medical Education 44(1), 85–93 (2010)
30. Vansteenkiste, M., Lens, W., Deci, E.L.: Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. Educational Psychologist 41(1), 19–31 (2006)
31. Wiener, E.L., Curry, R.E.: Flight-deck automation: Promises and problems. Ergonomics 23(10), 995–1011 (1980)

Designing the Interface to Encourage More Cognitive Processing

John Patrick[2], Phillip L. Morgan[1], Leyanne Tiley[2], Victoria Smy[2], and Helen Seeby[2]

[1]School of Psychology, Early Years, and Therapeutic Studies, University of South Wales, Caerleon Campus, Newport, NP18 3QT, UK
phillip.morgan2@southwales.ac.uk
[2]School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, UK
{patrickj,tileyl1,smyva,seebyh}@cardiff.ac.uk

Abstract. Cognitive engineering aims to provide operators with immediate access to as much relevant information as possible. However, this can encourage display-based strategies that do not involve committing information to memory. To overcome this problem, a somewhat counterintuitive method is discussed, based upon the theory of soft constraints [1], that involves delaying access to some critical information by one or two seconds. This design technique induces a more planful and memory-based strategy that can improve recall, develop more planning behavior, improve problem solving, and protect against the negative effects of interruption. Furthermore, we provide some preliminary results that this more memory-intensive strategy can be trained through past experience with high access cost and then used in situations where access cost is minimal. This was the case when only half of the training trials involved a higher access cost. Further research is needed to ascertain how long training effects last and what are the ideal training regimes for different types of task.

Keywords: Soft constraints, information access cost, strategy, memory, planning, problem solving, interruption, transfer.

1 Rationale for Increasing Access Cost

Interface design involves optimising interactions between human operators and the systems with which they work. Given increased technology, operators of complex systems can become deluged with information that they do not deeply process and adopt what is known as a 'display-based strategy', using the display as a form of external memory [1-3]. The unfortunate consequence is that critical information may not be processed deeply and subsequent performance may be impaired [3]. A novel, exciting, and counterintuitive solution, based upon both theory and empirical evidence, involves inserting a couple of seconds delay when operators attempt to access important information. Paying this small extra time cost induces a deeper cognitive processing strategy, involving more memory and planning, which improves performance in

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 255–264, 2014.

some task situations where such factors are important [3-6]. This paper discusses the theory underpinning this technique together with empirical evidence concerning its beneficial effect and how it may be used for training a more cognitively intensive strategy.

Imposing a small access cost encourages a shift to a more memory-based strategy by changing the cost/benefit balance facing the operator, making a display-based strategy less attractive. The theory underpinning this approach comes from Anderson's [7] seminal work on adaptive cognition and the theory of soft constraints [1]. Gray et al. [1] proposed that a task is made up of 'hard' and 'soft' constraints. Hard constraints dictate what behaviours are possible whilst soft constraints concern what strategy the operator chooses. The interface designer can manipulate the cost/benefit balance facing the operator by imposing a small cost in accessing information (a hard constraint) that will increase the degree of memory-based strategy (a soft constraint) adopted by the operator. This strategy shift occurs because participants find it beneficial to encode and plan what to do with the information rather than paying the access cost of a brief time delay on each occasion the information is viewed [1]. Therefore, somewhat counter-intuitively, increasing the cost associated with accessing information in an interface can induce the deployment of a more memory-based strategy that involves greater planning. A similar finding is that when the cost of executing a move is increased during problem solving, by requiring a series of extra key presses, this also results in increased planning of problem solving [8-9]. However these studies were concerned with the extra time cost associated with making a move rather than that concerned with accessing information, as in the studies reported in this paper.

2 The Effects of Inducing a More Memory-Based Strategy by Increasing Access Cost

Information access cost is defined as the time, physical and mental effort involved in accessing task critical information [1]. In studies examining its effect, access cost typically varies among three levels [3-4]. With a Low access cost, some important goal-state information is permanently visible at the interface. With a Medium access cost, this information is covered with a mask that disappears immediately when a mouse cursor is moved into the area containing the information and reappears when the cursor is moved out of that area. A High access cost has the same mouse movement as a Medium access cost although it also involves an extra second or two lockout time for the mask to disappear when attempting to view the goal-state information.

The result of increasing access cost is that the person engages in a more memory-based strategy, involving more encoding and planning. This was demonstrated originally by Fu and Gray [2] and subsequently in other studies [3-6] using the Blocks World Task (BWT, Figure 1). The BWT involves recreating a target pattern of colored blocks in a target window (top left of Figure 1) in a workspace window (top right of Figure 1) by dragging and dropping colored blocks from a resource palette (bottom of Figure 1). In the Waldron et al. [3] study, the three levels of Low, Medium

and High access cost were used between participants with a lockout time of 1 second in the High access cost condition and participants had to recall the goal pattern in a surprise recall test. There was a switch from a display-based strategy in the Low access condition to a more memory-based strategy in the higher cost conditions, with more blocks correctly recalled in the Medium and High cost conditions (Table 1).

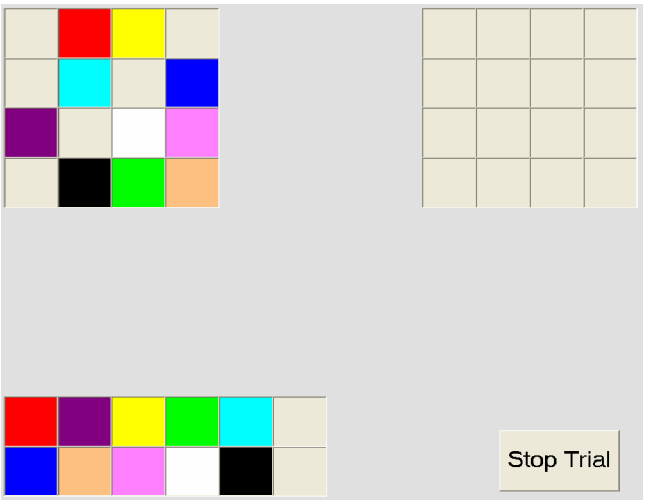


Fig. 1. Example of a Low information access cost start-state

Table 1. Effect of goal-state access cost on recall in the BWT [3: Experiment 1]

High Access Cost		Medium Access Cost		Low Access Cost	
Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
9.00	1.78	9.00	4.37	4.70	1.95

Typically better recall and better planning go together. A study by Waldron, Patrick and Duggan [10] directly investigated the effect of varying access cost on both the nature of planning (memory- versus display-based) and when planning occurred (before or during action) during problem solving. In a High access condition (with a 2.5 second lockout cost) more planning before action (i.e., before executing initial moves) was observed and less planning during action. This was demonstrated through longer first-move latencies and more moves executed per goal-state inspection. These findings therefore suggest that interface designers have a potential technique for inducing more intensive cognitive processing of information that involves greater encoding of information and planning of future moves.

However, this method is not a panacea as in some situations (e.g., fast-paced safety-critical environments) it may not be practicable or feasible and thus we need more evidence concerning when increasing access cost may be beneficial. Also, whilst one strategy for improving performance is to design the interface so that access cost is increased for some critical information, another strategy is to attempt to train

operators to adopt a more memory-based strategy for future situations when an access cost cannot or should not be included within an interface. Consequently the remainder of this paper considers two issues. First, what other situation(s) may benefit from a more memory-based planning strategy? Second, to what extent is it possible to train people, using the access cost method, to adopt a more cognitively intensive strategy?

3 Increasing Information Access Cost to Mitigate Negative Effects of Interruption

One practical situation that may benefit from an increased access cost approach is where performance is interrupted. Interruptions are intrinsic to our everyday lives and a wealth of applied and laboratory research evidence suggests that their effect is almost universally negative [13-16]. This has been demonstrated in various settings including offices, the flight deck, nuclear power plants and hospitals [17-18]. Negative effects include delays in resuming the interrupted task [13] [19], increased time to complete the interrupted task [20], and decreased accuracy of performance [21]. These performance deficits are likely due to the forgetting of task related goals during interruption [22] and one might expect such forgetting to be mitigated if a more memory-based processing strategy is developed before interruption.

Two studies are reported that investigated this issue. The first study by Morgan et al. [4], using the BWT, found that High access cost (involving a 2.5 second delay) reduced forgetting of planned copying moves following interruption, particularly when interruption occurred on half of all trials. It also improved prospective memory following two different interrupting tasks, even when one required the same type of processing resource as the primary task. A second study by Morgan and Patrick [5] examined whether higher access cost could protect against interruption during problem solving. Specifically, whether it was possible to induce more internal planning in the four disk Tower of Hanoi (ToH) that would result in not only more efficient problem solving but also increased resistance to interruption. In Experiment 1, more memory-based planning was developed by imposing a High access cost (with a 2.5 second delay to uncover the goal-state) that resulted in fewer moves to solution and the development of a more efficient sub-goaling strategy with more perfect solutions. In Experiment 2, High access cost protected performance against a ten second interruption irrespective of the interrupting task (blank screen-control, dissimilar mental arithmetic, or similar three disk ToH). Participants resumed more problems from memory with a High access cost and executed more moves after interruption without reviewing the goal-state (Table 2). Also fewer moves were required to complete interrupted ToH problems (Table 2).

Table 2. Effect of goal-state access cost on performance after interruption [5: Experiment 2)

Post-interruption performance measure	IAC	Mean	SD
Number of trials resumed without re-visiting goal state (max = 9)	High	6.30	1.86
	Medium	1.89	1.41
Number of moves executed without re-visiting goal state	High	5.33	2.24
	Medium	0.88	0.81
Number of moves to complete primary task following interruption	High	9.87	0.39
	Medium	10.20	0.35
	Low	11.27	0.44

The more memory-based planning strategy, induced by High access cost, presumably strengthened participants' goals during planning and problem solving, making them less susceptible to decay and interference from interruption. The finding that planning is enhanced due to higher access cost is important because planning rarely occurs spontaneously [11] and people are reluctant to use internal as opposed to external memory resources to plan [12]. Increasing access cost provides a means of encouraging more planning and protecting against forgetting following interruption.

4 Increasing Information Access Cost to Maximize the Utility of an Interruption Lag

Another proposed method for overcoming interruption effects is to use an interruption lag, which is a time delay between interruption annunciation (e.g., telephone ringing) and initiation of the interrupting task (e.g., answering telephone), during which the person can prepare to return to the interrupted task [23]. This provides an opportunity to rehearse a goal prior to task interruption and associate it with a salient reminder/priming cue. An interruption lag can lead to faster primary task resumption [13] although the time saved can be less than that imposed by the lag [16]. It is therefore desirable that the interruption lag is used as effectively as possible to encode goals prior to interruption. One means of trying to effect this is to introduce an access cost to induce a more memory-based strategy that should, in turn, result in better encoding of goals during the interruption lag.

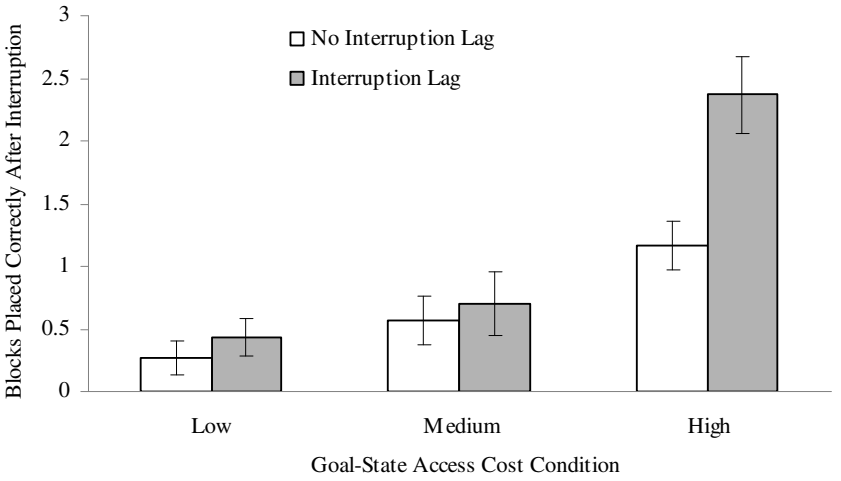


Fig. 2. Effect of access cost and interruption lag on recall after interruption [6: Experiment 2)

In order to examine this, Morgan et al. [6] used the BWT, and tested whether a 5-second lag was sufficient to recall planned moves and whether any benefit was dependent upon the strength of the memory-based strategy used to perform the task. Prospective memory was very poor with and without an interruption lag when the task was performed under Low and Medium access costs (Figure 2). However, prospective memory was not only improved under High access cost *without* an interruption lag, but this improvement was substantial *with* a lag (Figure 2).

5 Training a More Memory-Based Planning Strategy

The above studies have manipulated access cost through interface design and this has affected the degree of memory-based strategy adopted to perform the task. A further important issue is whether a high memory-based strategy can be trained and maintained when this access cost is no longer present in the design of the interface. The question is therefore whether it is possible to train a more memory-based strategy and the extent that this will carry-over to performance situations that do not involve an extra lockout time. This is not only a practical but also an important theoretical issue. The theory of soft constraints [1] emphasizes how the degree of memory-based strategy utilized to perform a task adapts to millisecond changes in the constraints of the current task environment. Gray et al. [2, p. 463] acknowledge that cost-benefit considerations may be overridden by factors such as training or by deliberately adopted top-down strategies. Interestingly, training relies on using past rather than current experience and therefore the following study investigated the relative contribution of past experience to present performance.

The study involved the BWT and a simple training and transfer design. Participants received twenty training trials with varying occurrences of high access cost (with a 2.5 second lockout cost in the High access cost condition) and then ten transfer trials under Medium access cost (i.e., with no lockout time). The extent that the high memory-based strategy, developed in training, was deployed in the medium access transfer environment is indicative of the degree of control exerted by past experience in the training environment as opposed to the constraints of the current transfer environment. It was predicted that increasing the amount of high access cost experienced during training would increase the degree of memory-based strategy deployed in the transfer environment.

5.1 Method

Participants. One-hundred and forty Cardiff University students participated for course credit and were randomly assigned to one of five conditions. There were 121 females and 19 males with an age range of 18-25 years (M=20.02 SD= 2.56).

Design. The experiment used the BWT and a between participants design with a training and transfer phase. The percentage of High access cost training trials was manipulated across five conditions ranging between 0 to 100% High access cost over 20 training trials (0%High = 20 Medium access cost training trials with no High access cost trials; 25%High; 50%High; 75%High; and 100%High). Medium access cost was used in the remaining training trials. Thus, in the 25%High condition, five of the twenty training trials involved High access cost and the other 15 involved Medium access cost. In order to facilitate comparison amongst these conditions on the training trials, five Medium access trials were presented on trials 1, 5, 10, 15 and 20 for all conditions, excluding 100%High. Apart from these trials, the order of High access cost trials was randomized. The transfer environment involved 10 Medium access cost trials.

5.2 Results and Discussion

One key measure of the degree of memory-based strategy is the number of blocks correctly copied after the first target window visit, which increases with more encoding and planning. The results are displayed in Figure 3 and it appears that more training with higher access cost led to a greater degree of memory-based strategy in both training and transfer situations. A 5 (condition: 0%High, 25%High, 50%High, 75%High, & 100%High) x 2 (environment: training and transfer) mixed ANOVA found, besides both main effects being significant, an interaction between the two variables, $F(4, 135) = 3.05$, $MSE = .26$, $p < .05$, $f = .30$. Participants in the 100%, 75%, and 50%High conditions copied more blocks correctly on the first visit to the target window than those in the 0%High condition in both training ($ps < .001$, .001, and .01 respectively) and transfer environments ($ps < .05$, .01, and .05 respectively).

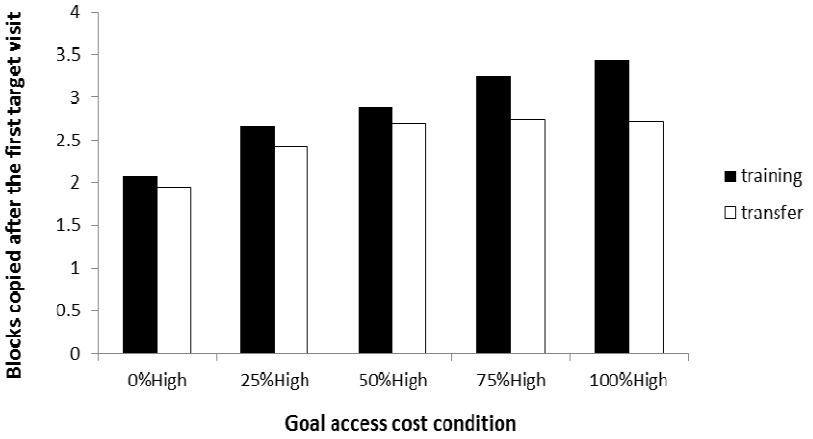


Fig. 3. Number of blocks copied after the first target window visit across training and transfer

Therefore more training with a high access cost increased the degree of memory-based strategy deployed in both training and transfer, as predicted. This indicates that there is an increasing carry-over effect from the training to the transfer trials because participants in the higher access cost training conditions did not exhibit the same degree of strategy as those without such training (Figure 3). Consequently we can conclude that the degree of memory-based strategy used depends not only on the current environment but also on past training. The effects of such training do not disappear immediately but are still evident over a series of 10 trials in which the access cost does not involve any lockout time.

This finding is important from both practical and theoretical perspectives. The results indicate that it is possible to increase the degree of memory-based strategy by not only providing 20 training trials with high access cost but also by dispersing fewer high access cost training trials amongst trials without such high access cost. The next research question is what is the minimal amount of training with high access cost necessary to achieve such an effect? From the current data, at least 50% of training trials are needed to increase the degree of strategy adopted when a lockout delay is no longer required to view the target pattern. However the current data does not inform us whether using less training trials than 20, and all with a high access cost, would also have a carry-over effect on transfer trials with no such cost.

6 Conclusions

Information access costs are not just an academic issue as they are intrinsic to our everyday computer environments when we try to avoid information clutter and when we have to pay such costs in opening and reopening applications, emails, or

documents. Access costs are also imposed when a password has to be recalled or the "terms" of an agreement have to be read. However, even though such costs occur naturally within computer environments, deliberately manipulating them may seem to contradict the traditional principles of cognitive engineering that strive to provide users with immediately relevant information that takes advantage of human perceptual abilities. For example, ecological interface design proposes that complex relationships between variables should be made directly accessible to operators in a manner that allows effortless extraction of information from the interface [24]. However, the theory of soft constraints [1] together with associated empirical evidence [4-6] suggest that this may not always be the best goal for interface design. We do not propose this as a panacea and practical applications have to be selected carefully. Imposing an access cost is a possible design strategy when important information has to be retained, possibly because it is no longer available or has to be acted upon in a timely manner without having time to search again for the information source. Such a situation is relevant not only to everyday computer-based work situations but also safety-critical environments in which incidents occur because of a failure to recall critical task information, sometimes with devastating consequences [25]. This paper has focused on two research areas that may benefit from the introduction of information access delays. First, it provides a new means of mitigating negative effects of forgetting following interruption by encouraging a more memory-based strategy with increased information access cost. It also facilitates the use of an interruption lag. Second, we have collected some preliminary evidence that it may be possible to train a more memory-based strategy although at present we need to know how long training effects last and what are the ideal training regimes for different types of task.

References

1. Gray, W.D., Simms, C.R., Fu, W.-T., Schoelles, M.J.: The soft constraints hypothesis. A rational analysis approach to resource allocation for interactive behavior. Psychological Review 113, 461–482 (2006)
2. Fu, W.-T., Gray, W.D.: Memory versus perceptual-motor tradeoffs in a blocks world task. In: Proceedings of the 22nd Annual Conference of the Cognitive Science Society, pp. 154–159. Erlbaum, Hillsdale (2000)
3. Waldron, S.M., Patrick, J., Morgan, P.L., King, S.L.: Influencing cognitive strategy by manipulating information access costs. The Computer Journal 50(6), 694–702 (2007)
4. Morgan, P.L., Patrick, J., Waldron, S.M., King, S.L., Patrick, T.: Improving memory after interruption: Exploiting soft constraints and manipulating information access cost. Journal of Experimental Psychology: Applied 15(4), 291–306 (2009)
5. Morgan, P.L., Patrick, J.: Paying the price works: Increasing goal access cost improves problem solving and mitigates the effect of interruption. Quarterly Journal of Experimental Psychology 66(1), 160–178 (2013)
6. Morgan, P.L., Patrick, J., Tiley, L.: Improving the effectiveness of an interruption lag by inducing a memory-based strategy. Acta Psychologica 142(1), 87–95 (2013)
7. Anderson, J.R.: The adaptive character of thought. Erlbaum, Hillsdale (1990)
8. O'Hara, K.P., Payne, S.J.: The effects of operator implementation cost on planfulness of problem solving and learning. Cognitive Psychology 35, 34–70 (1998)

9. O'Hara, K.P., Payne, S.J.: Planning and the user interface: The effects of lockout time and error recovery cost. International Journal of Human-Computer Studies 50, 41–59 (1999)
10. Waldron, S.M., Patrick, J., Duggan, G.B.: The influence of goal-state access cost on planning during problem solving. Quarterly Journal of Experimental Psychology 64(3), 485–503 (2011)
11. Delaney, P.F., Ericsson, A.K., Knowles, M.A.: Immediate and sustained effects of planning in a problem-solving task. Journal of Experimental Psychology: Learning, Memory, and Cognition 30, 1219–1234 (2004)
12. Hayes-Roth, B., Hayes-Roth, F.: A cognitive model of planning. Cognitive Science 3, 275–310 (1979)
13. Hodgetts, H.M., Jones, D.M.: Contextual cues aid recovery from interruption: The role of associative activation. Journal of Experimental Psychology: Learning, Memory and Cognition 32(5), 1120–1132 (2006a)
14. Hodgetts, H.M., Jones, D.M.: Interruption of the Tower of London task: Support for a goal-activation approach. Journal of Experimental Psychology: General 135(1), 103–115 (2006b)
15. Monk, C.M., Trafton, J.G., Boehm-Davis, D.A.: The effect of interruption duration and demand on resuming suspended goals. Journal of Experimental Psychology: Applied 14, 299–313 (2008)
16. Trafton, J.G., Altmann, E.M., Brock, D.P., Mintz, F.E.: Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. International Journal of Human-Computer Studies 58, 583–603 (2003)
17. McFarlane, D.C., Latorella, K.A.: The scope and importance of human interruption in human-computer interaction design. Human-Computer Interaction 17, 1–61 (2002)
18. Trafton, J.G., Monk, C.M.: Task interruptions. In: Boehm-Davis, D.A. (ed.) Reviews of Human Factors and Ergonomics, vol. 3, pp. 111–126. Human Factors & Ergonomics Society (2008)
19. Altmann, E.M., Trafton, J.G.: Timecourse of recovery from task interruption: Data and a model. Psychonomic Bulletin & Review 14, 1079–1084 (2007)
20. Eyrolle, H., Cellier, J.M.: Some effects of interruptions in work activity: Field and laboratory results. Applied Ergonomics 31, 537–543 (2000)
21. Flynn, E.A., Barker, K.N., Gibson, J.T., Pearson, R.E., Berger, B.A., Smith, L.A.: Impact of interruptions and distractions on dispensing errors in an ambulatory care pharmacy. American Journal of Health System Pharmacy 56, 1319–1325 (1999)
22. McDaniel, M.A., Einstein, G.O., Graham, T., Rall, E.: Delaying execution of intentions: Overcoming the cost of interruptions. Applied Cognitive Psychology 18(5), 533–547 (2004)
23. Altmann, E.M., Trafton, G.J.: Memory for goals: An activation-based model. Cognitive Science 26, 39–83 (2002)
24. Vicente, K.J., Rasmussen, J.: Ecological interface design: Theoretical foundations. IEEE Transactions on Systems, Man, and Cybernetics, SMC-22, 589–606 (1992)
25. National Transportation Safety Board: Aircraft Accident Report: Northwest Airlines, McDonnell Douglas DC-9-82, N312RC, Detroit Metropolitan Wayne Co., Airport, Romulus MI, August 16, 1987. Washington, D.C.: National Transportation Safety Board (1988)

The Measurement of Perceived Quality of Various Audio Sampling Rate and Frame Loss Rate

Xiangang Qin

Quality &User Experience, Mobile Business Group, Lenovo, P.R. China
jean.qxg@gmail.com

Abstract. In this paper, the influence of Audio Sampling Rate (ASR) and Frame Loss Rate (FLR) on perceived Quality of Experience (QoE) was studied. The result indicated that users are very sensitive to the damaged auditory quality caused by frame loss at 8 kHz and 12 kHz no matter how much it losses. The perceived damage of auditory quality caused by frame loss at 16 kHz and 24 kHz is also much lower that at 8 kHz and 12 kHz. Users even failed to perceive the negative impact of frame loss on auditory quality at 32 kHz whatever the frame loss rate is. The interaction effect indicates that users are not so sensitive to the negative impact of frame loss when the sampling rates increase to 16 kHz or higher.

Keywords: Perceived Quality of Experience, Audio Sampling Rate, Frame Loss Rate.

1 Introduction

Digital audio is the fundamental media in modern digital life. Currently almost all of the musical files preloaded in Smart Phone, Tablet and Smart TV, downloaded from internet or played online are digital audios. The advancement from analog audio to digital audio has significantly reduced the costs and improved the efficiency of distribution [1].

In digital audio system, sound is passed through an analog-to-digital converter (ADC) that converts an analog signal to a digital signal. The ADC runs at a specified sampling rate and converts at a known bit resolution [2].

The sampling rate defines the number of samples per second taken from a continuous (analog) signal to make a discrete (digital) signal [3]. The range of hearing for a healthy young person is 20 to 20,000 hertz [4]. According to Nyquist–Shannon sampling theorem, perfect reconstruction of a signal is possible when the sampling rate is greater than twice the maximum rate of the signal being sampled, or equivalently. The 44.1 kHz sampling rate used for Compact Disc was chosen for this and other technical reasons [3].

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 265–271, 2014.

Although having a sampling frequency more than twice the desired system bandwidth is desirable in some cases that extreme audio quality is required, lower sampling rates have the benefit of smaller data size and easier storage and transport [3]. That means the benefits of higher and lower sampling rates should be balanced in producing and designing the audios in a real and commercial system.

Audios are massively used in designing the interactions with users in Smart TV, such as the boot-up music, user manual, audio menu and controls and preloaded musical files, etc. The ideal audios should provide both good hearing experience that requires high sampling rate and good performance experience that requires small data size and quick transport.

Previous researches showed that most adults can't hear much above 16 kHz[5]. However, what will happen if frame is lost during transporting? Is 16kHz still the ideal sampling rate in balancing the data size and perceived quality of auditory experience.

This paper is dedicated to address this unanswered question.

2 Methodologies

2.1 Testing Stimuli

A audio file was original recorded in Chinese Language with Audacity(a audio editing software) with the sampling rate of 96kHz. It simulated a clip of typical dialogue "Xiaolin, Jin Wan You Kong Mei? Zan Lia Yi Qi Chi Ge Fan Bei' which means "XiaoLin, Are you free to have dinner with me this evening?" in English. The original audio file was then transformed into testing stimuli with various sampling rates ranging from 8kHz, 12kHz, 16kHz, 24kHz to 32kHz. The testing stimuli with various sampling were further transformed with various frame loss rate ranging from 0%, 1%, 3% to 5%. Totally, 20 (5 levels of sampling rates plus 4 levels of frame loss rates) testing stimuli were designed as the testing stimuli.

2.2 Testing Environment and Devices

The testing was conducted in a meeting room which simulated a typical living-room environment where the Smart TV was usually placed. The background noise is roughly about 45db. The audio files were played by a Lab-Top the speaker parameters of which are similar to that of Smart TV. The playing sound is about 78db. Users were seated in a chair which is about 1m away from the Lab-top

2.3 Participants

28 participants aged from 22 to 38 were invited to participate in the testing, half is Male and the other half is female. All of the participants have self-reported normal hearing ability.

Testing Procedure

The testing was conducted in four phases.

Warm-up Phase: The five audio files with various sampling rates were played in sequence from 8 kHz to 32 kHz for users to get the baseline of rating the perceived quality of experience.

1st Testing Phase: Mean Opinion Score (MOS) of sampling rates. In this phase, MOS is used to rate the perceived Quality of Experience of the five audio files with various sampling rates respectively in random. A rating scale from 1-5 was used to rate the perceived experience level of the speaking, in which 1 represents Bad and 5 represents Good. Each audio file with different sampling rate was played one time and users can ask the moderator to play the same audio file once again if it is needed.

2nd Testing Phase: Pair Comparison of Sampling Rates. In order to explore the possibility of differentiating the perceived experience between pair of sampling rates, Pair Comparison Method (PCM) was used in this phase. 10 pairs of audio files with various sampling rates were compared in random sequence. After each pair of audio files was played, user would orally tell which one is better in terms of the perceived quality of experience.

$$C_5^2 = \frac{5}{2 \times (5-2)} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = 10$$

3rd Testing Phase: Pair Comparison of Frame Loss. In this phase, PCM was used to evaluate the influence of various frame loss on perceived quality of experience of audio files with various sampling rates. However, the comparisons were only made among various frame loss within one sampling rate and the frame loss across different sampling rate were not compared. Altogether, 30 pairs of audio files with various frame loss were compared.

$$5 \times C_4^2 = \frac{4}{2 \times (4-2)} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} = 5 \times 6 = 30$$

3 Results

3.1 Mean Opinion Score of Sampling Rates

The result of MOS shows significant main effect of sampling rates, $F(4,112)=8.14, p<0.01$, The perceived QoE of 16Khz, 24Khz and 32Khz is significantly higher than that of 8khz and 12khz. No significant difference was found between 16khz, 24khz and 32khz. It indicates that perceived QoE of 16kHz reaches plateau and sampling rate which is higher than 16khz contributes little to the improvement of perceived QoE.

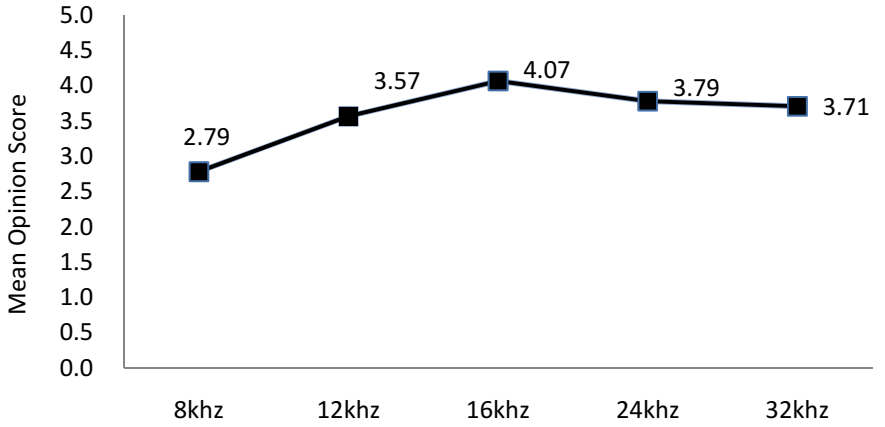


Fig. 1. The Mean Opinion Score of Various Sampling Rates

3.2 Pair Comparison of Sampling Rates

The result of Pair Comparison of Sampling Rates shows that the probability of being perceived better than a lower sampling rate declines when the sampling rate is 16kHz or higher. This result indicates that the perception of the differences between high sampling rates gets difficult.

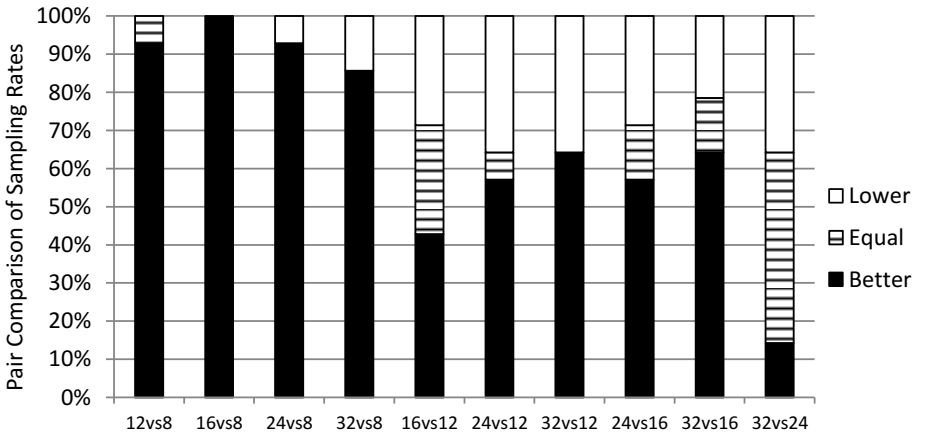


Fig. 2. Pair Comparison of Sampling Rates

3.3 Pair Comparison of Frame Loss

The result of PCM of Frame loss reveals significant main effect in the perceived QoE of six pairs (0% vs 1%, 0% vs 3%, 0% vs 5%,1% vs 3%,1% vs 5%,3% vs 5%) of various frame loss rate. *F(5,120)=38.57,p<0.01*.The perceived QoE of loss rate at 1%, 3% and 5% is significantly lower than that at 0%. The corresponding damaging value is -0.64 , -0.75 and -0.64 respectively on the -2 to +2 rating scale in which – means the perceived QoE is damaged and + means it is improved. However, no significant difference is found between loss rate at 3% and 5%.

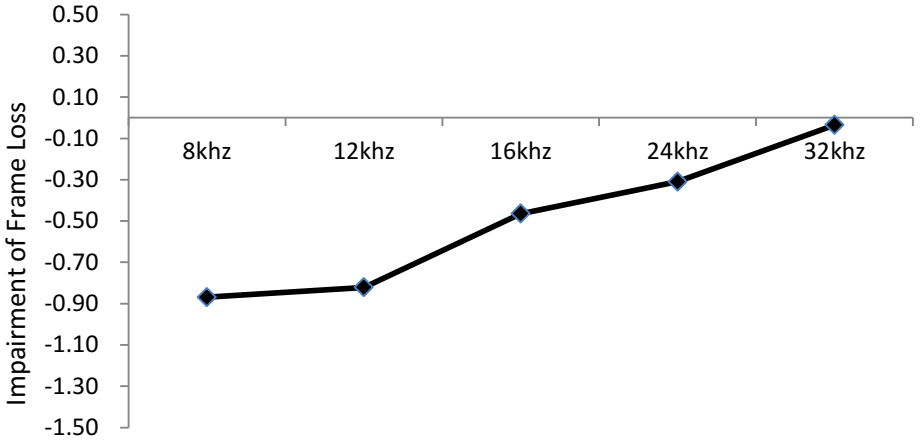


Fig. 3. Impairment of Frame Loss on Perceived Quality of Experience with Various Sampling Rates

Significant interaction effect is found in sampling rate and frame loss rate , F(20,540)=9.33,p<0.01. When the sampling rate is at 8khz and 12khz, users perceived significant damage of auditory quality at 1%, 3% and 5% frame loss rate in comparison with 0%. However, the perceived damage of auditory quality is much lower when comparison is made between various frame loss rates. The result means that users are very sensitive to the damaged auditory quality caused by frame loss at 8khz and 12khz no matter how much it losses. The perceived damage of auditory quality caused by frame loss at 16khz and 24khz is also much lower that at 8khz and 12khz. Users even failed to perceive the negative impact of frame loss on auditory quality at 32khz whatever the frame loss rate is. The interaction effect indicates that users are not so sensitive to the negative impact of frame loss when the sampling rates increase to 16khz or higher.

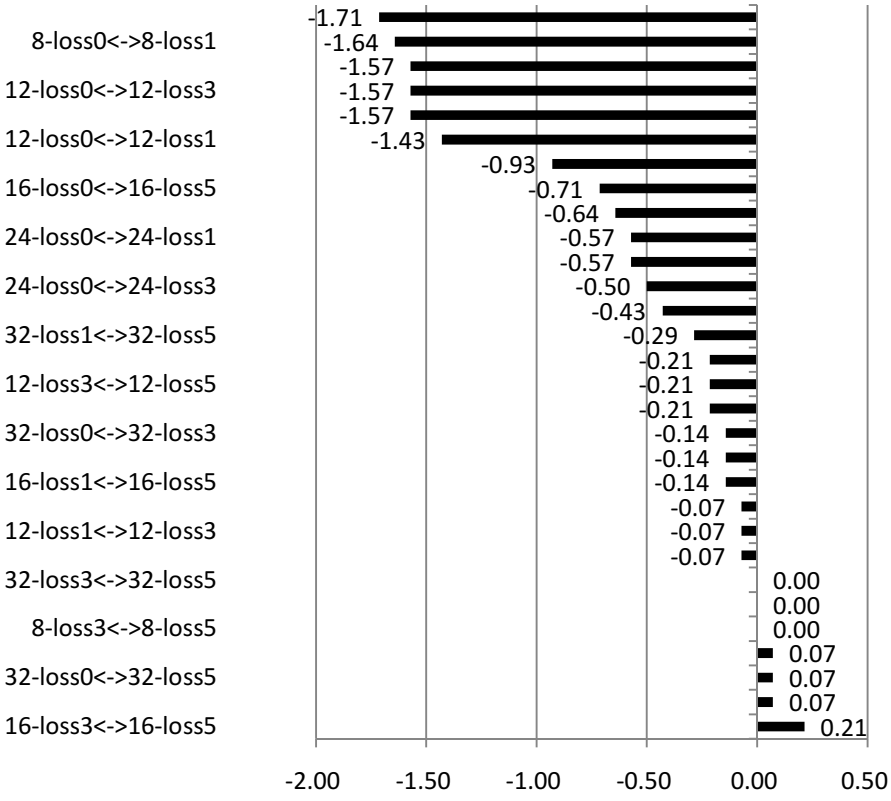


Fig. 4. The Comparison of Perceived Quality of Experience of Various Frame Loss

4 Discussion

4.1 The Measurement of Perceived Quality of Experience of Audio File

In this study, Mean Opinion Score and Pair Comparison Method were used as the evaluating methodologies. Although MOS measures the independent perception of the experience and PCM measures the dependent perception of experience which relies heavily on the relative differences between audio files, this study found consistent results in the findings by different measurement methods. It validates the reliability of this study.

4.2 16 kHz is the Turning Point of Perceived Quality of Experience of Sampling Rates

The result of MOS shows that the perceived experience increases sharply with higher sampling rates but reaches a plateau when it is 16 kHz and higher. The negative impact of frame loss on perceived QoE also decreases sharply when the sampling rate is 16 kHz or higher. This means 16 kHz is a "golden" sampling rate which makes desirable balance between perceived QoE and efficiency of data transportation.

4.3 Frame Loss Impaired the Perceived QoE Heavily with Low Sampling Rates

When audio files with low sampling rates (lower than 16 kHz) were used in designing the auditory interface of products, the insurance of data transportation is key to the satisfaction. Even only 1% frame is lost that users could clearly perceive the impairment of QoE. Audio files with 16kHz or higher is preferable.

4.4 Balance between Sampling Rate and Frame Loss

Although it's desirable to use high sampling rate in designing the auditory elements of HCI system, the redundancy of higher sampling rate shouldn't be wasted because the ability of most ordinary humans is unable to distinguish the supposed advantages of higher sampling rate over low sampling rate when it's higher than 16 kHz. The redundant resources should be used to improve the benefits of small data size, data transportation, etc.

5 Summary

Based on the results of this study, audio files with sampling rate of 16 kHz should be used in designing the auditory interaction of Smart TV which makes desirable balance between perceived QoE and data transportation.

References

1. Janssens, J., Vandaele, S., Beken, T.V.: The Music Industry on (the) Line? Surviving Music Piracy in a Digital Era. European Journal of Crime 77(96) (2009)
2. Wikipedia , http://en.wikipedia.org/wiki/Digital_audio
3. Wikipedia, http://en.wikipedia.org/wiki/Sampling_rate
4. Frequency Range of Human Hearing,
5. http://hypertextbook.com/facts/2003/ChrisDAmbrose.shtml
6. Shepherd, I.: Are high sample rates making your music sound worse?, http://productionadvice.co.uk/high-sample-rates-make-your-music-sound-worse/

Defining and Structuring the Dimensions of User Experience with Interactive Products

Jean-Marc Robert

Polytechnique Montréal, Department of Mathematics and Industrial Engineering
P.O. 6079, St. Centre-Ville, Montreal, Quebec, Canada H3C 3A7
jean-marc.robert@polymtl.ca

Abstract. The goal of this research is to define the dimensions of User Experience (UX) with interactive products and systems in order to lay the ground for the construction of a subjective assessment tool for UX. After defining UX, we describe several characteristics of UX and present key elements of some UX models in order to understand the ins and outs and the process of UX. Then we present the results of two empirical studies wherein 77 persons were asked to tell UX stories with products. From their stories we extracted 12 UX dimensions which can be grouped around two poles : Product and User. Thereafter we present the underpinning model and an outline of a new UX subjective assessment tool based on the assessment model of NASA-TLX, a well-known tool for assessing mental workload. As conclusion, we indicate the next steps of the construction and validation of the new tool.

Keywords: User Experience (UX), UX dimensions, Interactive product, Subjective assessment, Assessment tool, NASA-TLX.

1 Introduction

Since the beginning of the year 2000, the concept of User Experience (UX) has gradually dethroned (without rejecting) the concept of usability to account for the quality of our interactions with different products, systems, or services [N.B. : for brievity, we only use the word product in the rest of paper]. Even a large professional association, called UPA, changed its name for UXPA (User Experience Professional Association) in 2012 to mark the turning.

The concept of UX was rapidly adopted by the communities of Industrial Design, Interaction Design, Human-Computer Interaction, and Ergonomics/Human Factors which are all concerned by the quality of products and the challenge of creating positive UX with them. Not surprisingly, UX was accepted from the start for products that are intended for a large public, such as Web sites, smart phones, video games, and popular software applications where the issues of pleasure, emotions, and aesthetics are important, and where there is a fierce competition on the market. But UX is more and more adopted by designers of serious products which are clearly associated with work, performance, efficiency, and security, such as aircrafts [1], flight simulators, production planning software, schedule optimizer software, etc. Because positive UX

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 272–283, 2014.

has a great customer appeal and that counts on the market. The reasons of this rapid and widespread adoption are numerous: UX is considered as a richer, more global, more inclusive, more interesting, and more profitable concept than usability. It is directly related to the goal that is being pursued : create a rich and positive experience that will help to sell the product. It is clearly another mean to satisfy the customers' claims and expectations, to project a better corporate image, and to be more competitive on the market.

Despite its short history, UX has already been the topic of an abundant scientific literature [7] [18] so that several definitions, models, methods and tools for approaching UX are available. Yet the evaluation of such a phenomenon remains a challenge because it is subjective, multidimensional, dynamic, and context-dependant. As for usability three major categories of measures are available: physiological (in rapid progress), behavioral, and subjective. In this paper we will focus on subjective evaluation because it is rich and lends itself well to the capture of complex subjective phenomena. Our goal is to identify and define UX dimensions and sub-dimensions, and use them in new subjective assessment tool for UX. We will present an outline of the tool we are developing.

This paper is structured as follows : after this introduction, first we present some definitions and different characteristics of UX; second we describe the key elements of some UX models in order to understand the ins and outs and the process of UX; third, we define several UX dimensions and show data on their frequence and importance; fourth, we present the assessment model of the NASA-TLX (Task-Load Index), a well-known tool for evaluating mental workload which can serve as model for UX evaluation; and finally we present an outline of a new UX evaluation tool. In the conclusion, we suggest some actvities for constructing and validating the tool.

2 UX Definitions

We selected three definitions of UX from a large set of definitions in order to show different facets of UX.

First, the ISO 9241 definition [5] states that UX is "*A person's perceptions and responses that result from the use and/or anticipated use of a product, system or service*." So UX is a *subjective* phenomenon. What is open to criticism here is that an UX could result only from the anticipated use of a product. While fully recognizing the role of anticipation in the creation of an UX, it is difficult to imagine an UX without a real use of the product.

Nielsen-Norman's definition [11] states that « *User experience" encompasses all aspects of the end-user's interaction with the company, its services, and its products*". So UX is not only concerned with the product itself but also with the company and its services. « All aspects » include the different activities with or about the product, such as searching for information, buying, downloading, installing, learning how to use, using, repairing, doing the maintenance, dealing with the customer service, installing the uddates, etc. So UX is *global* and *cumulative.*

Finally, Robert & Lesage's definition [15] states that « *UX is a multidimensional construct that defines the overall effect over time on the user of interacting with a*

system or service in a specific context ». Several characteristics of UX stand out : apart from being a construct, it is multidimensional, it is an overall and cumulative effect that builds up with time, and it is situated in a context.

3 UX Characteristics

To have an UX you must have a User who interacts with a Product for doing an Activity in a some Context. These basic elements are present in several UX models ([2] [4] [7] [15] [17]). If there is no product in use (e.g., looking at a sunrise), there is no UX. Different types of products may be involved: interactive (e.g., smart phone, video game), adjustable (e.g., mountain bike, car seat), a combination of both, or not adjustable (e.g., seat in the subway). Finally the user of a product may be active (e.g., using the Wii console), creative (e.g., designing), or passive (e.g., sitting in a train). In this paper, we focus on interactive products and active users.

In light of several UX definitions and UX stories ([15]), we can identify several characteristics of UX:

- *subjective*: "UX happens inside the person" ([17]); since UX is based on user's perception and responses, it is a personal and subjective. Several characteristics of the person are involved: knowledge, abilities, goal, motivation, philosophy, past experience, values, attitudes, expectations, preferences, sensitivity to aesthetics, anxiety, fatigue, culture [8], etc.
- *multidimensional*: When persons report their usage of different products and tell the reasons why they appreciate or criticize them, several factors emerge: it is easy, they get the right information, they feel competent, they exchange with people, they have fun, they feel cool, etc. These terms correspond to different UX dimensions.
- *holistic*: UX covers all aspects of our interactions with the company, its services and its products. This includes all the steps we go through with the product when we search information, buy, transport, install, learn how to use, use, talk to the customer service, etc. UX is the global result of our perceptions and responses at each step.
- *situated in a context*: Elements of context are very diversified: time pressure, period of the day, location, presence and pressure of people, weather conditions, competition, issues at stake, etc. They definitely impact on UX and contribute to determine its positive or negative valence, its strength, its memorable character.
- *dynamic*: UX evolves in time as we develop abilities with the product, go through positive and negative experiences, operate in easy/difficult personal conditions (e.g., stress, fatigue), and test different contexts. So UX becomes richer, more complete, more precise, better defined. Each new usage is likely to alter our perceptions.
- *cumulative*: This quality is the corollary of the previous one. UX depends both on the our expectations about the product based on what we saw (publicity), heard, read or imagined about it, on the real experience of using it, and on the global evaluation we make afterwards, when we combine the past and the present.

– *several granularity levels*: UX can be about the interaction with a single product for doing an activity in a short period of time (e.g., register at a terminal) as well as the interaction with several products for doing several activities over a long period of time (e.g., planning and doing a flight trip).

4 4UX Models

Several UX models ([4] [9] [10] [15]) present important elements that come into play to create UX. Fortunately they have much in common.

First, as seen above, to have a UX, there mut be an *User interacting* with a *Product* for doing an *Activity* in a *Context*. These elements are indispensable.

Second, when we interact with a product, we come into contact with its *instrumental* and/or *non-instrumental* qualities. The former refer to what the product enables us to do (*do-goals*) and to the external services it provides: for instance, make a phone call, take a photo, compose a letter, send an email, etc. The latter refers to what the product enables us to be (*be-goals*) and to the inner satisfaction it brings to us because of weelbeing, self-achievement, sense of aesthetics, etc. Fortunately these two sets of qualities are not exclusive of each other. Instrumental qualities are more naturally associated with *extrinsic goal*s, i.e. that are external to us, like making money, having good marks, winning a contest, etc. And non-instrumental qualities are naturally associated with *intrinsic goals*, i.e., that are internal to us, like having fun, feeling good, achieving oneself, etc. When we use the product (and after using it), we have various perceptions, feelings, and emotions which create our UX with the product and make it positive/negative, rich/poor, striking/not striking, memorable/not memorable.

Third, the outcomes of UX may different forms. If the UX is positive, we will be encline to reuse the product, buy it, talk about it positively, recommend it to others, do something to support it, etc. Furthermore, our attitudes and expectations will be positive for the next usage of the product.

5 UX Dimensions

What we call an UX dimension is a major or significant factor that can explain the creation of an UX. Based on the results of empirical studies presented below, we distinguish two types of dimensions: those that are the Product input to the UX creation and those that are the User input to the UX creation. Let us illustrate the difference with two examples. When a mother says "I like my cell phone because it allows me to always keep in touch with my daughter", it is the Usefulness dimension of the product that stands out and contributes to create the UX. When an user says "I like playing with this video game because I have fun", it is the Psychological dimension (subdimensions: pleasure) that stands out. It is essential to identify and define UX dimensions rigorously because they form the basis of the new tool we are developing for assessing UX. Several studies allowed us to identify and define UX dimensions, either directly or through the UX models they present.

Robert & Lesage [15, 16] interviewed and/or observed six persons interacting each with a different system (smart phone, Wii console, mountain bike, interactive

monitor in airplane, …) for doing different activities (work, communicate, play video games, go biking, …) in different contexts (at home, in transport, in a city, in the wild). With this material they could construct six UX stories of 10-15 lines each describing what the user is doing and what his/her UX consists of. Here is an excerpt of the UX story of the mountain biker: " … He specifically enjoys those short challenging segments that require all his attention, physical abilities and wits, for minutes on end; and lead him to total, exhilarating exhaustion after two hours or so. The usability of the bike, although essential, is obviously just a part of the interaction with the device. UX is more a question of extreme fun, strong emotions, hard challenge, pride, intense physical effort, acquisition of abilities, and self-accomplishment". They extracted six dimensions of UX: Functional, Physical, Perceptual, Cognitive, Psychological, and Social. They will be defined in the next section.

The research work of Larouche [6] and Robert & Larouche [14] aimed at identifying, defining, and measuring the frequency and importance of UX dimensions in order to lay the ground for the construction of a subjective evaluation tool. Based on a literature review Larouche identified nine UX dimensions : the six ones mentioned above plus *Contextual, Informational*, *Cultural* [8]. She then collected empirical data to test the frequency and importance of these dimensions. To do so, she asked 52 persons in a questionnaire to describe a positive and a negative UX with two different products of theuir choice they had interacted with. Three participants reported only one UX story so that there was a total of 101 UX stories. The products involved in the UX stories were very diversified : 72 interactive products (Web sites, smart phones, video games, …), 12 adjustable products (ergonomic chair, car seat for children, bike, …), 10 interactive and adjustable products (video camera, Global Positioning System, …), and 7 not adjustable products (train, bus, subway, …). Three judges with good knowledge in Ergonomics/ Human Factors for Human-Computer Interaction examined the UX stories, individually at first and then collectively, in order to: a) check if the nine dimensions were present in the UX stories and measure their frequency; b) evaluate their importance on a Likert scale (0 = absent, 1 = slightly, 2 = moderately; 3 = very) ; the average of the three scores of judges was calculated for each dimension; c) search for new UX dimensions; and d) see if the same dimensions appeared in positive and negative UX stories. Results indicate the presence of nine UX dimensions : eight of the above ones and temporal (see Larouche in Table 1); they will be defined in the next section. *Temporal* includes the time saved or lost. Some dimensions (*Functional, Psychological, Cognitive, Contextual, Informational*) are much more frequent and important than others. The same dimensions appear in positive and in negative UX: so no dimensions are exclusively related to negative or positive UX. We are critical about the *Contextual* dimension because we rather see it as one of the four basic components of any UX. For this reason, we consider it should not be considered as a dimension.

Provost [12] and Provost & Robert [13] pursued the same objectives as Larouche : identify, define and measure the frequency and importane of UX dimensions; Provost used a different method and appealed to a different group of persons to collect empirical data on UX. In light of a literature review, she identified 10 UX dimensions. To test them, she conducted semi-structured interviews by phone with 25 persons, asking them to tell a positive and a negative UX stories with a product of

Table 1. Dimensions and sub-dimensions extracted from UX stories

Robert & Lesage [15]	**Larouche** [6] *	**Provost** [12] **		
Dimensions	Dimensions		Dimensions	Sub-dimensions
Functional	Functional 96%; i: 2,32	Pole Product	Functionality 88%; s : 4,25	usefulness, functionality options, possibilities
			Usability 88%	simplicity, rapidity ease of use, efficiency
	Informational 74%; i: 1,48		Informational 70%; s: 3,68	presence, relevance, quality
Physical	Physical 50%; i: 0,81		Physical characteristics 42%; s: 2,88	weight, dimensions size, adjust-ments
			External characteristics 56%	customer service brand, eco-system
			Other qualities 48%	accessibility, secur-ity, reliability avail-ability, robustness
	Contextual 79%; i: 1,46			
	Temporal 49%; i : 0,77			
Perceptual	Perceptual 54%; i: 0,85	Pole User	Perceptual 66%; s: 3,18	aesthetics, presence & quality of multi- media, sense stimulation
Cognitive	Cognitive 80%; i: 1,68		Cognitive 74%; s : 3,34	understanding, concentration, learning reflection, attention memory, stimulation
Psychologi-cal	Psycholo-gical 90%; i: 1,92		Psycholo-gical 90%; s: 3,68	pleasure /frustration, motivation, expectations, values, evocation, meaning; positive emotions: negative emotions
Social	Social 49%; i: 0,73		Social 54%; s: 1,80	presence of others, quality of inte-ractions in/dependence from/ to others, ob-taining info about others
			Physical 40%	physical activity, transport, com-fort movement, displacement
			Other person-al impacts 62%	productivity profitability return on investment

* The % is the ratio: Nb of UX stories wherein the dimension is present / 101 UX stories;
i (importance) is the average of the three judges' scores (0=absent, 1=slightly, 2=moderately, 3=very).

** The % is the ratio : Nb of UX stories wherein the dimension is present / 52 UX stories.

s (strength) indicates the evaluation of the strenght of each dimension made by the user on a 5- point scale (0: nul, 1: very low; 2: low; 3: moderate; 4: high; 5 : very high). The user answers to this question : To what extent do you think the dimension has contributed to your UX? Some results sre not available.

their choice, explain the reasons why it was positive and negative, and complete an evaluation grid about the UX dimensions present in the story (the grid was given to the participant after s/he had told his/her two stories). The interviews were recorded to allow three judges to listen to them and extract UX dimensions. The products mentioned in the UX were very diversified : web sites, software and personal computers, small electronic devices (camera, video camera, …), transportation (cars, motorcycles, bikes, …), etc. Results indicate that 12 dimensions can account for the positive and negative UX stories with a large variety of interactive systems (see Provost in Table 1). Interestingly, these dimensions can be grouped around two poles : Product and User. The pole *Product* encompasses six dimensions (the % indicates the frequency of the dimension in the 52 UX stories) : Functionality (88%), Usability (88%), Informational (70%), Physical characteristics (42%), External characteristics (56%), other Qualities of the product (48%). The pole *User* also encompasses six dimensions : Perceptual (66%), Cognitive (74%), Psychological (90%), Social (54%), Physical (40%), et other Personal impacts (62%). The number of dimensions present in each UX varies from one UX to another. Results confirm Larouche's findings, showing that the same dimensions can be found in positive and negative UX. Finally the study allowed to find several sub-dimensions; these will be useful to orient the construction of the evaluation tool.

6 Definitions of UX Dimensions

In light of the above results, in the following paragraphs we define eight basic UX dimensions and their sub-dimensions (see Table 1).

Functional : This dimension corresponds to qualities that make a product reliable, compatible with others, accessible, available, and well adapted to its physical and human environment.

- Reliable: quality of a product that works without failure, that does not break easily when it hits something or when dropped.
- Compatible: quality of a product that is well integrated with its environment, its ecosystem, and that can therefore be used in conjunction with other products.
- Accessible: quality of a product that meets the needs of specific users: for example the disabled, elderly, people with reading deficiencies, in disabling conditions.
- Available: quality of a product that can be ased at any time or when users need it, and in any place or where users need it.

Usefulness/Usability

- Usefulness: quality of a product that enables the user to satisfy his/her needs and achieve his/her objectives.
- Usability: quality of a product that is easy to learn and use. The ISO definition is more elaborate : it also includes efficiency and user's satisfaction.
- Performance characteristics: these include for example response speed, memory capacity, computing power, and image quality.

Informational: This dimension corresponds to the utility, right balance, and appropriateness of the information provided by the product depending on the context. It includes two sub-dimensions:

- Quality of information : the product provides information that is reliable, exact, precise, and accessible both in its form and its content.
- Quantity of information : the product provides exhaustive information with a degree of finesse and precision sufficient for the user in a given context.

Physical characteristics (under the pole Product):.

- Physical characteristics: include for example the weight, the shape, the dimensions (e.g., keyboard, display), the battery life, ...

Sensory/Perceptual: This dimension corresponds to the impression left by the product on the sense organs, to the impact on the user's perception. It includes three sub-dimensions:

- Visual : all that is related to the appearance and aesthetics of the product and that is perceived by the user.
- Hearing : all that is related to the sound emitted by the product during its use and that is perceived by the user.
- Tactile : all that is related to tactile sensations with the product and that is perceived by the user.

Cognitive: This dimension refers to human information processing done while using the product; it includes different types of activities such as analyzing, evaluating, reflecting, learning, creating, etc.

- Cognitive effort /development: the use of the product sollicits or stimulates the user's intellect to acquire new knowledge or skills, analyse, reflect, solve problems, respond quickly, etc.

Psychological : This dimension refers to the emotions felt by the user when s/he interacts with the product, and to the values and opinions that this interaction triggers. It includes several sub-dimensions:

- Stress: the use of the product generates stress to the user because of an imbalance between what is required to interact with the product and the resources or time that are available.
- Pride: the use or possession of the product brings a high sense of dignity, honor or satisfaction to the user.
- Pleasure: the use or possession of the product brings a state of contentment to the user because s/he satisfies a need or a desire.
- Frustration: the product brings dissatisfaction to the user, for example because it is blocked or it does not meet the user's expectations and desires appropriately.
- Evocation: the product evokes memories for the user.

- Attachment: this corresponds to the intensity of emotional attachment to the product.
- Moral value: the use or possession of the product reflects life principles that guide moral judgment of users and apply to their consciousness as an ideal.

Social : This dimension is about linking the user with other people through the product. It includes two sub-dimensions:

- -Contact: the product allows the user to contact and interact with other people.
- -Culture: the product allows to connect the user to his/her culture which is defined by the set of distinctive spiritual, material, intellectual and emotional features that characterize a society or social group.

Physical (under the pole User)

- Physical effort/stimulation: the use of the product requires exertion or physical activity from the user.

Synthesis
The underlying rationale of trying to evaluate UX only with these dimensions and sub-dimensions is the following. Although it is obvious that the user's predispositions, the context of use, and the types of user's activities are likely to have a determinant impact on UX, we do not take them explicitely into account when evaluating UX because we do no control them. We rather focus on the user's perception of the product characteristics (dimensions around the pole *Product*) because we control these characteristics as designers, and we focus on the user's appreciation, feelings and emotions (under the pole *User*) because they are at the core of UX. The perception of the dimensions under the pole *Product* is the perception of the input stimuli to UX whereas the perception of the dimensions under the pole *User* is the perception of the user's responses to these inputs. We suggest to add *Aesthetics* as a dimension under the pole *Product*, because it is clearly a product characteristic, it is not fully covered by the dimension Physical characteristics, it has become an important issue of several products, and it is likely to impact the user's perception. The Physical dimension appears on both sides but with different meanings. Under the Pole *Product*, it means that the user perceives the objective product characteristics (e.g., life battery, weight) whereas under the pole *User* it means that the user gives an appreciation of the physical effort or stimulation.

7 NASA-TLX as an Assessment Model for UX

There are strong similarities between the concepts of UX and mental workload: both belong to the domain of Cognitive Ergonomics and are subjective, multidimensional (with a limited number of dimensions), and holistic. The results of their evaluation depend, among other things, on both the user's predispositions, the type of activity, and the context. Of course there also exists a major difference between the two: mental workload cannot be negative whereas UX can be so. Yet it is worth verifying if

lessons can be learned from the way mental workload is evaluated with the NASA-TLX tool, a widely accepted tool in the community [3].

NASA-TLX is a questionnaire completed by the person who accomplishes a task. It includes six dimensions of workload : Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration Level. The first three dimensions deal with the evaluation of work demand such as it is perceived by the user, whereas the last three ones deal with the evaluation of the person's reactions. The tool allows the person to indicate which dimension is present or not for a task and weigh it, and it provides a global score of workload. Here is the procedure of use :

- The user performs a task; during that task or right after it, s/he completes the NASA-TLX;
- The person evaluates each of the six dimensions on a 20-point Likert scale;
- The person circles the name of a dimension on each of 15 cards showing different pairs of dimensions; then we count the number of times each dimension was chosen and this gives the weight of each dimension;
- The results of evaluations in the previous two steps are multiplied together : this give the adjusted score for each dimension;
- the total of the adjusted scores is divided by 15 : this gives the global score of mental workload.

8 Building an UX Assessment Tool on the Model of NASA-TLX

In our opinion a similar approach to that of NASA-TLX can be followed for UX assessment. The procedure will be as follows :

- ask the user to use an interactive product for doing an activity in a certain context, or ask him/her to think about a past interaction with a product; during the interaction or after it, use the UX evaluation tool;
- ask the user to indicate on a Likert scale if each dimension (see Table 1) is present or not in his/her UX with the product;
- ask the user to weigh each dimension; here we cannot use the same stratagem as NASA-TLX with its 15 cards showing two dimensions among which we must choose. The reason is that with 12 dimensions or so, the number of pairs of dimensions explode so that it would be much too long for the user to circle one dimension per pair;
- combine the two evaluations to calculate the adjusted score of each dimension;
- add the adjusted scores: this gives the total of adjusted scores;
- divide this total by the number of dimensions and ths gives a global score of UX.
- The tool offers these facilities:
- - it provides the user with definitions of UX dimensions and sub-dimensions to facilitate the use of the tool;
- - it allows the user to indicate if the UX is positive, neutral or negative;
- - it allows the user to revise an evaluation already entered;

- - it calculates automatically the adjusted scores of each dimension and the global score;
- - it gives an overview of the dimensions and sub-dimensions that come into play in an UX.

9 Conclusion

In this paper we defined several UX dimensions and sub-dimensions for the assessment of UX with interactive products. These dimensions form two groups which can be placed under two poles : Product and User. We used the NASA-TLX, a well-established tool for the subjective assessment of mental workload, as a model of assessment for UX, because of strong similarities between the two concepts of UX and mental workload. A very promising idea taken from NASA-TLX for UX assessment is to allow the user to give a double evaluation of each UX dimension : the first is about the presence of a dimension in the UX, and the second is about the weigh (or importance) of this dimension on the UX. We described the rationale and procedure of UX assessment following the model of the NASA-TLX. We gave an outline of the UX assessmen tool. The next step of this research will consist in prototyping and testing the tool, validating it with UX experts and end-users, and evaluating how it really helps designers and developers to improve their products.

References

1. Ahmadpour, N., Lindgaard, G., Robert, J.-M., Pownall, B.: The Thematic Structure of Passenger Comfort Experience and its Relationship to the Context Features in the Aircraft Cabin. Ergonomics (to appear, 2014)
2. Arhippainen, L., Tähti, M.: Empirical Evaluation of User Experience in Two Adaptive Mobile Application Prototypes. In: Second International Conference on Mobile and Ubiquitous Multimedia, Norrköping, Sweden, pp. 27–34 (2003)
3. 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)
4. Hassenzahl, M.: The Thing and I: Understanding the Relationship between User and Product. In: Blythe, M., Overbeeke, C., Monk, A.F., Wright, P.C. (eds.) Funology: From Usability to Enjoyment, pp. 31–42. Kluwer Academic Publishers, Dordrecht (2003)
5. International Organization for Standardization: Ergonomics of Human System Interaction - Part 210: Human-centred Design for Interactive Systems. International Organization for Standardization, ISO 9241-210 (2008)
6. Larouche, A.: Survey for Determining UX Dimensions. Report for a Master's degree, Polytechnique Montréal, 92 pages (2011) (in French) (unpublished)
7. Law, E., Vermeeren, A., Hassenzahl, M., Blythe, M. (eds.): Towards a UX Manifesto: COST294-MAUSE Affiliated Workshop. Lancaster, UK (2007)

8. Lee, I., Choi, G.W., Kim, J., Kim, S., Lee, K., Kim, D., Han, M., Park, S.Y., An, Y.: Cultural Dimensions for User Experience: Cross-Country and Cross-Product Analysis of Users' Cultural Characteristics. In: Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction (BCS-HCI 2008), vol. 1, pp. 3–12. British Computer Society, Swinton (2008)
9. Mahlke, S.: User Experience: Usability, Aesthetics and Emotions in Human-Technology Interaction. In: Law, E., Vermeeren, A., Hassenzahl, M., Blythe, M. (eds.) Towards a UX Manifesto: COST294-MAUSE Affiliated Workshop, Lancaster, UK, pp. 26–30 (2007)
10. Mahlke, S., Thuring, M.: Usability, Aesthetics and Emotions in Human-Technology Interaction. International Journal of Psychology 42(4), 253–264 (2007)
11. Nielsen Norman group: Our Definition of User Experience, http://www.nngroup.com/about/userexperience.html (accessed November 8, 2009)
12. Provost, G.: Study of User Experience with Interactive Products. Report for a Master's degree, Polytechnique Montreal, 92 pages (2012) (in French) (unpublished)
13. Provost, G., Robert, J.-M.: The dimensions of Positive and Negative User Experiences with Interactive Products. In: Proceedings of HCII (Humnan Computer Interaction International), Las Vegas, Nevada, July 21-26, 10 pages (2013)
14. Robert, J.-M., Larouche, A.: The Dimensions of User Experience with Interactive Systems. In: Blashki, C. (ed.) Proceedings of IADIS International Conference - Interfaces and Human Computer Interaction, and Game and Entertainment Technologies, July 21-23, pp. 89–96. IADIS Press, Lisbon (2012)
15. Robert, J.-M., Lesage, A.: Designing and Evaluating User Experience. In: Boy, G.A. (ed.) Handbook of Human-Computer Interaction, pp. 321–338. Ashgate, U.K. (2011a)
16. Robert, J.-M., Lesage, A.: From usability to User Experience with Interactive Systems. In: Boy, G.A. (ed.) Handbook of Human-Computer Interaction, pp. 303–332. Ashgate, U.K. (2011b)
17. Roto, V.: User experience from product creation perspective. In: International Conference HCI 2007, Lancaster, UK, pp. 31–34 (2007)
18. Tullis, T., Albert, W.: Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Morgan Kaufmann, Burlington (2010)

Misperception Model-Based Analytic Method of Visual Interface Design Factors

Xiaoli Wu[1,2], Chengqi Xue[2], and Zhou Feng[1,2]

[1] College of Mechanical and Electrical Engineering, Hohai University, Changzhou 213022, China
wuxlhhu@163.com, yutakazf@gmail.com
[2] School of Mechanical Engineering, Southeast University, Nanjing 211189, China
ipd_xcq@seu.edu.cn

Abstract. The unreasonable design of interface information has given rise to malfunctions of cognition and decision-making among operators, thus leading users into a complex cognition and finally resulting in serious failures in information recognition and analysis, and even in operation and execution processes, which poses one of the major causes for many accidents. Firstly, there remains an internal relevance between errors and perception, and five error factors i.e., visual confined, visual interference, visual illusion, attention shift and over attention were extracted from the point of visual attention mechanism; Secondly, the cognitive model (theory) and psychological experimental paradigm corresponding to the cognitive level were combed out through explanation of the error level, thus the misperception model was established; Finally, It provides a feasible basis for design improvement of visual interface through behavior and physiological experimental data. This misperception analysis method of visual interface has applied mature psychological experimental paradigm and can favorably analyze the design factors from the aspect of misperception, so as to play a significant role in improving the visual interface design.

Keywords: Visual interface, Visual attention mechanism, Misperception model, Design factors, Error factors, Psychology experiment.

1 Introduction

With the rapid development of industrial design and computer interactive media, visual information interface has become an essential information interactive medium in a complex system. The unreasonable design of interface information has given rise to malfunctions of cognition and decision-making among operators, thus leading users into a complex cognition and finally resulting in serious failures in information recognition and analysis, and even in operation and execution processes, which poses one of the major causes for many accidents. Errors are common human failures occurring in information interface and its cognition mechanism of errors is an important hitting-point for improving interface design as well as the key for reducing

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 284–292, 2014.

cognition difficulties. In this paper, by extracting the error factors from the visual attention mechanism, the misperception model build a method to analyze the visual interface design factors, which provides a feasible basis for design improvement of visual interface through behavior and physiological experimental data.

2 Misperception Model

2.1 Extraction of Error Factors in Visual Attention Mechanism

The ability to quickly find what objects are interesting and meaningful out of large amounts of visual data is called visual selective attention[l-4]. This is an essential feature of visual information processing formed during the long and complicated evolution and development of biological visual system and its interaction with nature[5-6]. There are two types of visual attention processes: the first one is pre-attentive process which focuses on attract attention, the second one focused attentive process which is a process of pay attention[7]. What guides visual attention can be bottom-up data driving factors or top-down task driving factors. Accordingly, the selective attention of visual information can be realized through two patterns. First, the reason why an individual chooses a stimulus in the visual attentive field is that the individual considers the stimulus very important in reaching the current task objective. Therefore, under this condition, the task objective and subjective intention of the individual command the visual attentive process. Cognitive psychologists call this kind of active selective attention mechanism goal-directed attention or top-down selective attention; second, because of the fact that a stimulus within the visual field has extraordinary features different from other surrounding objects, the individual's attention is automactically captured by this conspicuous stimulus, in spite of the current task objective or subjective intention of the individual. This type of passive selective attention mechanism is called stimulus-driven attention or bottom-up selective attention, as shown in Fig.1.

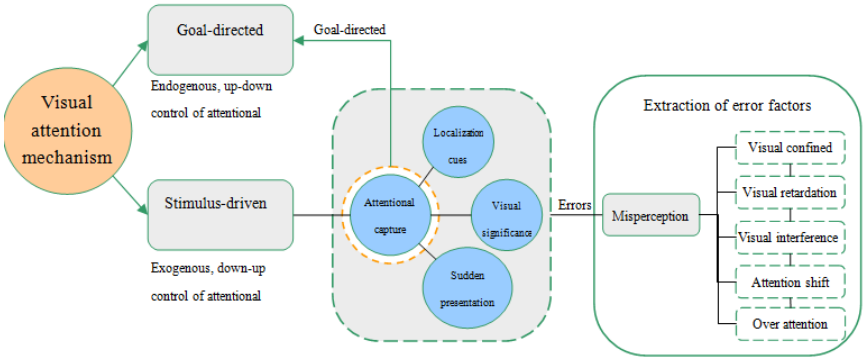


Fig. 1. Extraction of error factors in visual attention mechanism

From the error types research perspective, Norman [8] (1981) proposed to classify errors as three types, i.e., error, fault and failure. Reason [9-10] (1987) put forward eight basic error types including misperception and attention failure on the basis of the theory of Norman; Li Leshan [11] (2004) considers negligence and over attention as two primary aspects leading to users' errors. Supported by previous study results, i.e., error types of misperception proposed by Norman and Reason, this paper has further established the connection of errors and perception through visual attention mechanism. Misperception factors are extracted as visual confined, visual illusion, visual interference, attention shift and over attention respectively, as shown in Fig.1

2.2 Establishment of Misperception Model

This model is divided into error level and cognitive level; error level includes five types, i.e., visual confined, visual illusion, visual interference, attention shift and over attention. Through error explanation, the error level will shift to relevant cognitive model (theory) and psychological experimental paradigm corresponding to the cognitive level, as shown in Fig.2.

1. Visual confined: blind, visually impaired, or out of sight
2. Visual retardation: Reaction was too late, no stimulation or stimulation is not obvious
3. Visual interference: the target transfer, distracted because of the stimulus is not obvious
4. Attention shift: visually impaired, unable to cope with multiple targets
5. Over attention: overly concerned and do not know or misjudgment

That can be entering cognitive level. Through explanations of error level, we have obtained the corresponds of cognitive level with relative cognitive processing theories. Those are visual confined to visual search, visual retardation to perceptual organization, visual illusion to cognitive laziness & cognitive busyness, visual disturbance to attentional capture, attention shift to top-down attentional set, and over attention to focusing attention & sustained attention. We can establish cognitive stratified model of misperception, as shown in Figure 2. They will be explained from cognitive theories.

There are several relative theories in cognitive stratified model of misperception, such as preview search, perceptual load theory, prioritizing selection mechanism, schema model, cognitive load theory, perceptual selected model, susceptibility to interference model, energy distribution, biased-competition model, attentional load theory, and so on. Conceptually-matching theory, that can be called top-down processing, states subjective factors for the guiding role of the perceptual process. The factors are knowledge experience, motivation and expectation. The interpretation mechanism of prioritizing selection in preview search in detail and comprehensively verified the visual-marking-based interpretation in the two aspects: the location-based inhibition and feature-based inhibition of old distractors（Hao F., 2006）[12]. Visual prioritizing selection indicated that the visual system would give priority to the

selection of current behavioural and target-related stimulus and ignore irrelevant stimulus (Han Sh. H, 2000) [13]. Cognitive busyness or cognitive laziness and believed that the lack of motivation caused cognitive laziness and that high processing load caused cognitive busyness (Pett and Wogeber, 2001) [14]. Neisser proposed Attention model-schema model and believed that the attention was not a filter or attenuator. The event importance was not responsible for memory entering decision. And the objects of the individual attention were closely related to the task-activated schema (Neisser, 1976) [15]. Lavie proposed attention load theory and believed that attention selection depended on the amount of current processing resources, which had limited capacity (Lavie et al., 2004, 2005) [16-17]. The biased-competition model based on object attention, in which Desimone proposed that many objects in visual search scene competed attention resources for attention resources were limited. If the characterization of certain object is the same to the target template kept in the current working memory, the object will gain a competitive advantage in prioritizing selection for visual attention (Desimone et al., 1995) [18]. The susceptibility to interference model, in which Dempster proposed that the inhibition processing regulated the resistance to interference and that interference resistance efficiency

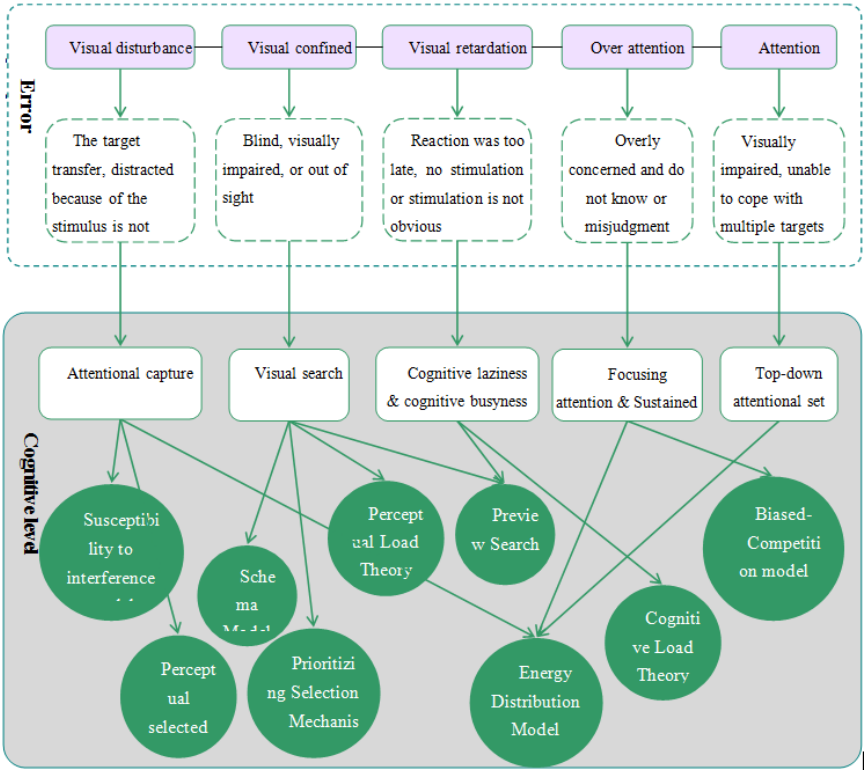


Fig. 2. Misperception Model

reflected the individual susceptibility to interference, the operation-consumed susceptibility under a variety of interference stimulus conditions (Dempster et al., 2003) [19]. These cognitive processing theories above provide base for the study on qualitative analysis experiments of misperception model.

3 Method

Based on the misperception model, a psychological experimental method has been built for analyzing visual interface design factors, which could be conducted as per the following three steps: firstly, extract the error factors of visual interface design factors to be analyzed; then, select relevant cognitive model (theory) and psychological experimental paradigm and apply the error factors as independent variables to design the experiment; lastly, employ the method combining reaction time and eye movement tracing to carry out the experiment. The analysis of variance method is adopted to statistically analyze the indexes of reaction time, and error rate as well as the indexes of saccade frequency and saccade amplitude in physiological data, thus obtaining the misperception analysis result of visual interface information factors, as shown in Fig.3.

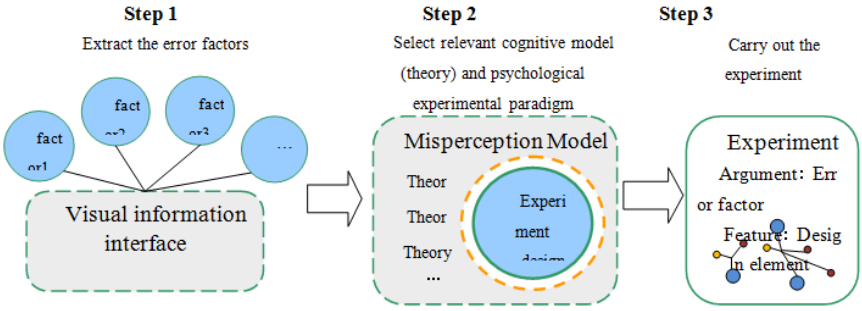


Fig. 3. Analytic method of design factors in visual interface

4 Example

4.1 Extract the Error Factors

In this paper, as an example of visual interface in complex system, the error factors of interface task are extracted. There are 15 tasks information in this radar situation-interface, and it can be listed 17 probable error factors from information display corresponding four monitoring tasks, which are surveillance/discover, status query, response plans and response execution. The text step is error characterization according to misperception model, which are visual confined, visual interference, visual retardation, attention shift and over attention. It can be selected error factors of

visual interference as the independent variables in order to carrying out our experiment with corresponding experimental paradigm.

4.2 Select Relevant Cognitive Model (Theory) and Psychological Experimental Paradigm

There are several relative theories and model for attentional capture in cognitive stratified model of visual interference, such as perceptual selected model, susceptibility to interference model, biased-competition model, attentional load theory, and so on. As for the experimental paradigm study on visual search, Theeuwes[20-21] et al. (1998, 2004) have held the opinion that, the occurrence of attention capture mainly depends on the significance level of the feature of one stimulus relative to that of other stimuli. The higher the feature significance level of a stimulus, the higher the possibility of its generating attention capture. Fleetwood and Byrne[22-23] (2002, 2006) have found through experimental observation that, the first factor which influence the user's visual search is the quantity of icons, the second is the target boundary, and the last one is the quality and resolution of icons. Patrick[24] (2003) has applied the experimental paradigm of visual delay search task to comparatively study the binding experiment of colors, positions as well as colors and positions, the results showed that the binding experiment had not obviously shortened the search time compared with the other two groups of experiments. Yu Bolin[25] et al. have studied the role of word gap played in visual interference, adopted the same-different matching task and visual search with the time series presented by word gap and stimulus as the variables. These two experiments validated that, word gap is a necessary and sufficient condition for visual interference derived from context. Van Orden[26] et al (1993) have employed the brightness and flash as the ways of highlighting to study the shapes and colors of symbols, and testified to the influence of symbol shapes and colors on search time. Wickens[27] et al. (1990) have studied the information identifications in different color codes and spacial positions under multiple information channels. According to the above experimental study review of visual search, information symbols such as color, shape, typeface, position as well as icon quantity and quality have possessed a large experimental study basis, and their interactive interference effects have been primarily demonstrated by experiments. Although these are all basic psychology experiments, they have certain reference value for practically applied of radar situation-interface, based on which, the preliminary presupposition of this experiment could be obtained.

4.3 Experiment

This paper is to conduct an experimental study on fighter situation-interface feature search: simulate the radar situation-interface of complex system, extract the associated factors of visual interference, apply the technological means of psychological experiment on such aspects like interference environment and features of information matter, carry out the visual interference experiment of target search, analyze the attentional capture and search strategy of identification of different

information matters in different interference environments according to the features of visual selective attention and explore the law of the influence of interference environment and information matter features on visual search.

4.4 Result

Based on previous studies, the information matter in the radar situation environment was designed into three kinds of different feature items as per the shape and color type features. Data indicated that (Fig.5, Fig.6), under low-interference and high-interference environments, with the progressive increase of quantity, feature items showed an obvious trend of progressive increase in reaction (Fig.5), and feature 3 (irregular shape-hybrid colored feature item) consumed the longest time; more reaction time was needed in high-interference environment than in low-interference environment. Data also indicated that (Fig.7, Fig.8), the error rate presented a law different from that in reaction; in low-interference environment, feature items 1, 2 and 3 showed a trend of progressive increase, which suggests that regular-shaped single colors are easier to search than irregular-shaped single colors and hybrid colors, and not susceptible to causing misjudgment issues; in high-interference environment, the error rate showed no obvious trend of progressive increase, which also suggests that, in an environment with multiple interfering objects.

Under both low-interference and high-interference environments, the reaction time and error rates of the subjects were tested when the three different feature items were presented by different quantities. Variance analysis on the reaction time indicated that, the main effect of feature items under low-interference environment ($F=24.781$, $P=0.001$, $p<0.05$) and the main effect of feature items under high-interference environment ($F=10.184$, $P=0.012$, $p<0.05$) both had reached the significance level (as indicated by Fig.4); variance analysis on the error rate indicated that, the main effect of feature items under low-interference environment ($F=5.297$, $P=0.047$, $p<0.05$) had reached the significance level, while the main effect of feature items under high-interference environment ($F=1.613$, $P=0.275$) was insignificant (as indicated by Fig.5).

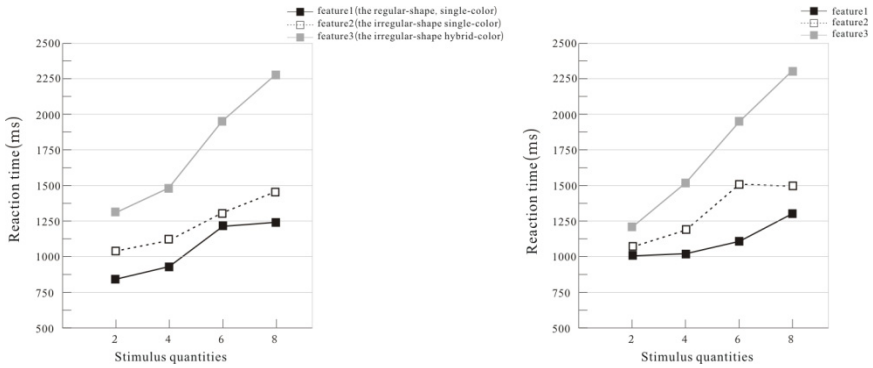


Fig. 4. Reaction time of the three feature items under low-interference and high–interference environment

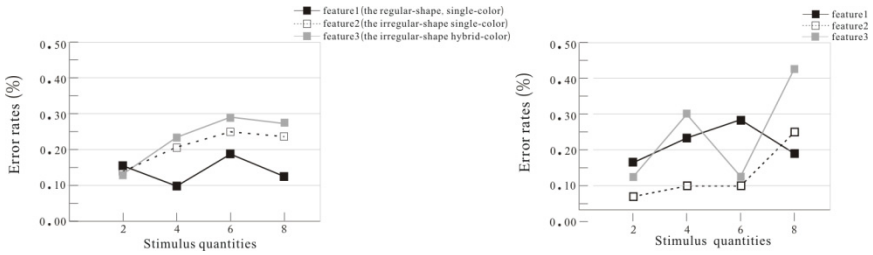


Fig. 5. Error rates of the three feature items under low-interference and high–interference environment

5 Conclusion

1. There remains an internal relevance between errors and perception, and five error factors i.e., visual confined, visual interference, visual illusion, attention shift and over attention were extracted from the point of visual attention mechanism;
2. The cognitive model (theory) and psychological experimental paradigm corresponding to the cognitive level were combed out through explanation of the error level, thus the misperception model was established;
3. Error factors were applied as independent variables to change previous psychological experimental method and were able to analyze the key factors of visual interface design on the aspect of misperception.
4. Interference environment and information matter features both have played important roles in influencing the information identification in radar situation-interface, which is the design factor needing to be considered in the information layout of complex situation-interface.

This misperception analysis method of visual interface has applied mature psychological experimental paradigm and can favorably analyze the design factors from the aspect of misperception, so as to play a significant role in improving the visual interface design.

Acknowledgment. This work was supported by the Social Science Fund for Young Scholar of the Ministry of Education of China(Grant No. 12YJC760092), Fundamental Research Funds for the Central Universities (Grant No. 2013B10214), the National Nature Science Foundation of China (Grant No.71071032,71271053).

References

1. Shen, J.: Guidance of eye movements during conjunctive visual search: The distractor-ratio effect. Dissertation Abstracts International: Section B: The Sciences and Engineering 63(2B), 6126 (2003)
2. Schweizer, K.: Visual search, reaction time, and cognitive a-bility. Perceptual and Motor Skills 86(1), 79–84 (1998)

3. Williams, C.C.: Age differences in visual memory and the re-lation to eye movements and executive control processes in visual search. Dissertation Abstracts International: Section B: The Sciences and Engineering 64(8B), 4080 (2004)
4. Underwood, G., Chapman, P., Brocklehurst, N., UnderwoodJ, C.D.: Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers. Ergonomics 46(6), 629–646 (2003)
5. Lee, T.S.: Computations in the early visual cortex. Journal of Physiology 97, 121–139 (2003)
6. Grill-Spector, K., Malach, R.: The human visual cortex. Annual Review of Neuroscience 27, 649–677 (2004)
7. Norman, D.A.: Categorisation of action slips. Psychology Review 88, 1–15 (1981)
8. Reason, J.: Human Error. Cambridge University Press, New York (1990)
9. Reason, J.: Human error: Models and management. British Medical Journal 320, 768–770 (2000)
10. Leshan, L.: Human computer interface design. Science Press, Beijing (2004) (in Chinese)
11. Fang, H.: Uninhibited processing mechanism of prioritizing selection in preview search. Institute of Psychology, Chinese Academy of Sciences, Beijing (2006)
12. Shihui, H.: The global precedence in visual information processing. Journal of Chinese Psychology 32(3), 337–347 (2000) (in Chinese)
13. Petty, R.E., Desteno, D., Rueher, D.D.: The Role of Affect in Attitude Change. In: Forgas, J.P. (ed.) Handbook of Affect and Social Cognition, pp. 212–229. Lawrence Erlbaurn, Mahwah (2001)
14. Neisser, U.: Cognition and reality: principles and implications of cognitive psychology. W.H.Freeman (1976)
15. Lavie, N., Hirst, A., de Fockert, J., Viding, E.: Load theory of selective attention and cognitive control. Journal of Experimental Psychology: General 133, 339–354 (2004)
16. Lavie, N.: Distracted and confused? Selective attention under load. Trends in Cognitive Sciences 9, 75–82 (2005)
17. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18, 193–222 (1995)
18. Bjorklund, D.F.: Children's thinking, 4th edn., 119–140, 142–145,162–163, 343–344. Wadsworth/Thomson Learning, Belmont (2003)
19. Theeuwes, J., Burger, R.: Attentional Control during Visual Search the Effect of Irrelevant Singletons. Journal of Experimental Psychology: Human Perception and Performance 24, 1342–1353 (1998)
20. Theeuwes, J.: Top-down Search Strategies Cannot Override Attentional Capture. Psychonomic Bulletin & Review 11(1), 65–70 (2004)
21. Fleetwood, M.D., Byrne, M.D.: Modeling icon search in ACT-R/PM. Cognitive Systems Research 3, 25–33 (2002)
22. Fleetwood, M.D., Byrne, M.D.: Modeling the Visual Search of Displays:A Revised ACT-R/PM Model of Icon Search Based on Eye-Tracking and Experimental Data. Human-Computer Interaction 21(2), 153–197 (2006)
23. Monnier, P.: Redundant coding assessed in a visual search task. Displays 24(1), 49–55 (2003)
24. Bolin, Y.: An experimental study of effects of character spacing under visual interference. Psychological Science 17(3), 129–132 (1994) (in chinese)
25. Van Orden, K.F., Divita, J., Shim, M.J.: Redundant use of luminance and flashing with shape and color as highlighting codes in symbolic displays. Human Factors 35, 195–204 (1993)
26. Wickens, C.D., Andre, A.D.: Proximity compatibility and information display: effects of color, space, and object display on information integration. Human Factors 32(1), 61–77 (1990)

Positive Affective Learning Improves Memory

Chen Yang[1, 2], Luyan Ji[1, 2], Wenfeng Chen [1], and Xiaolan Fu[1]

[1] State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
yanchen2460@163.com, {jily,chenwf,fuxl}@psych.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. It is well documented that affective learning materials can impact learning process, but it is unclear that what the role of affective learning-irrelevant stimuli is. To tackle this issue, this study provided evidence that affective value of irrelevant stimuli can be transferred to learning materials and influences the learning process. Using a variant of minimal affective learning paradigm, the experiment demonstrated that only one occasion of neutral-affective pairing can lead to affective learning, and showed an advantage of positive affective learning on the improvement of face memory. Implications for the design of affective human-computer interactive system are discussed.

Keywords: affective learning, affective interaction, face learning, memory.

1 Introduction

Information technology has reshaped learning and instruction, and more and more learning activities are undergoing with interactive tutorial systems. Research has shown that users tend to interact with the virtual agents in tutorial system in the same way as they would with humans [1]. Human-computer interaction could be improved if machines can naturally adapt to their users, and response in an adaptive manner according to the users' affective state as well as cognitive state. The research on affective computing have proved that it is important to interact with emotional information involved, possibly including expressions of frustration, confusion, disliking, interest, and more [2]. Affective computing expands human computer interaction by including emotional communication together with appropriate means of handling affective information [2].

Emotional interaction is essential for emotional intelligence [3], and humans assess emotional signals from themselves and/or others, with varying degrees of accuracy. The goal of affective computing is to give computers skills of emotional intelligence, including the ability to recognize and express emotion as a person might [4]. Researchers in educational technology field had make great effort to make the educational systems more customized for the affective states of the learners [5]. Emotional interaction is especially important for learning system, given that emotions are well known to play significant role in learning. During the learning activities, humans tend to generate affective experiences, and emotions permeate educational contexts and

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 293–300, 2014.

affect everyone in the learning process [6]. On the one hand, the user will feel a number of emotions during their interaction with this kind of system. For example, the user may enjoy the learning process when he/she thinks the system is helpful, or become frustrated when her/his expectations are not satisfied. On the other hand, emotional outcomes of learning will play a highly important role in learning performance. For example, negative emotions experienced by the user can affect their perception of the interface, and prevent them from concentrating on and remembering information [7], and, thus, lead to lower learning performance.

However, not only can the learning outcomes, but also the affective aspects of the system interface lead to affective processing, e.g., the virtual emotional behavior and affective characteristics of the agents. During social communication in reality, we pay attention to not only the biological characteristics of the communicators, such as one's age, gender, race and so on, and also social aspects of the persons, such as emotional state and attractiveness. These sorts of information can bias our social behaviors as well as cognitive process in social communication. For example, teachers may show different emotions ranging from anxiety to joy and pride in the classes, and both teaching and learning involve emotional understanding [8]. Thus, teaching is a form of emotional behaviors, and teachers' emotion may play an important role in interactions in classroom and learning performance [8]. By being impolite in online system, virtual agents are not viewed as technologically deficient, but they are more likely believed as humans by the users. Therefore, the virtual emotional behavior of the agents is also very important for the interaction during learning, and might have a significant impact on learning performance [9].

Although it has been evident for the relationship of virtual emotional behaviors and learning, it remains unclear whether the affective characteristics of the agents also impact learning. In a recent study, Plass et al. (2014) provided evidence in support that the affective design of multimedia learning materials can be used to foster positive emotions, and such positive emotions can facilitate learning [10]. They found that round face-like shapes both alone and in conjunction with warm color induced positive emotions; and comprehension task was facilitated by warm colors and round face-like shapes, independently as well as together, and transfer learning was facilitated by round face-like shapes when used with neutral colors. Magner et al. (2014) found the affective interface design, e.g., learning-irrelevant decorative illustrations, can foster learning for near transfer for those students with high prior knowledge [11]. It was suggested that affective decorative illustrations might foster situational interest and, thus, ease the learner's focusing of attention and reducing effort of cognitive activation. These results suggest that affective characteristics of the learning system interface can impact learning. Thus, it is reasonable to postulate that affective characteristics of virtual agents can also impact learning.

Nevertheless, irrelevant affective decorative illustrations can also hinder learning (e.g., near transfer) [11]. To interpret this fact, it was suggested that decorative illustrations can be considered as seductive details that require investment of cognitive resources that might then not be available for processing essential information. Mayer et al. (2005) suggested that cognitive resources during learning may be available as a result of optimized design of the learning environment [12]. This is a cognitive

account, but not affective account. Thus, although it is evident that affective value of learning materials can impact learning process [10], it is unknown that whether this is the case for affective values of learning-irrelevant stimuli (e.g., decorative illustrations).

It has been well documented that affective stimuli are more efficiently encoded, consolidated, and retrieved than neutral stimuli [13-14]. In this line of study, memory for faces is a crucial topic for psychology and social science since the processing and the encoding of emotional signals conveyed by faces is used to form impressions, to evaluate the intentions of others and to adapt future behavior [15]. The affective characteristic of face (e.g., facial expression and attractiveness) is also proved to play a key role in affective and social behavior as well as facial encoding and recognition memory.

Apart from the affective characteristics conveyed by face, the individual behavior description [16] can be also affective. Those affective personal descriptions and faces integrate together will become a unify characterization in memory. For example, neutral face will be impressed as more attractive if it has been associated with positive trait (e.g. honest) [17]. However, it is unknown whether face memory is influenced by paired irrelevant affective information during learning phase.

Research on affective learning has accumulated evidence that affective values of learning-irrelevant stimuli can be transferred to neutral learning materials. That is, when the neutral learning contents are accompanied with emotional context, they may acquire affective value via affective learning process [18]. There is accumulating evidence that people can learn the affective value of individuals from detailed behavioral descriptions of those targets. However, it remains unclear how the transferred affective value influence learning process. The present study was to tackle this issue and investigated the mechanism of affective learning effect on memory using simple affective stimuli. In this study, face learning task was used as an analogous task to simulate the role of learning-irrelevant affective characteristics of virtual agents in learning. We used face encoding as learning task, and face recognition memory as learning performance.

2 Method

2.1 Participants

Eleven undergraduates (6 males and 5 females, the average age is 21.9 years (SD = 2.1)) participated in this experiment for a small payment. All had normal or corrected-to-normal vision. Participants in the experiment were naive about the purpose of the study.

2.2 Stimuli

Forty photographs of Chinese faces taken by the researchers were used in the face learning task (20 for learning targets and 20 for distractors). Half of the photos were male faces. Each photo shows a close-up full face with a neutral expression. The face

covers approximately 80% of the photo, posed against a plain light-colored background. The size of each photo is 220×300 pixels and results in a visual angle of 5.3°×12.6° at the distance of about 60 cm. Each photo was assigned a unique name, occupation, and residence (i.e., a city/town and a state/province of participants' residence). The names were randomly selected from telephone books. All names contain three characters.

2.3 Design and Procedure

It was a one-factor design with affective valence (positive, negative) of the trait descriptions as within-participant variable.

A variant of the minimal affective learning paradigm (Bliss-Moreau et al., 2008) was introduced. The stimuli were presented on a 17'' computer monitor with an E-prime program. The experiment consisted three phases: learning phase, filling phase, memory test phase and affective evaluation phase. Four practice trials were provided before the actual encoding task to familiarize participants with the task.

During the learning phase, participants viewed 20 sequentially presented face-sentence pairs for 5 s and were told to learn and remember the pictures. Each of the twenty target faces was paired with a unique descriptive sentence that was positive or negative in affective tone. The descriptions are sentences introducing the name, the occupation, the residence of the person in each photo, and whether s/he is a good/evil person. Each trial started with a centered fixation cross for 1000 ms, then replaced by a photo presented for 5000 ms, along with a simultaneously presented description below it. A 1000-ms blank screen was presented as an inter-stimulus interval before proceeding to the next trial.

During the filling phase, participants completed some unrelated arithmetic tasks for 5 min to prevent rehearsal.

During the memory test phase, the 20 target faces in learning phase and 20 novel faces were sequentially presented. Each trial started with a centered fixation cross for 1000 ms, followed by a photograph. The photo was terminated by a key press response before moving to the next trial. Participants were asked to judge whether the face was old one in the learning phase.

During affective evaluation phase, the 20 target faces in learning phase and 20 novel faces were sequentially presented again. Participants were told to make two judgments: valence categorization task and liking rating task. Each trial started with a centered fixation cross for 1000 ms, followed by a photograph. Participants were asked to judge whether the face was positive, neutral, or negative and how s/he likes the picture based on a 5 points scale.

3 Result and Discussion

3.1 The Affective Learning Effect

The valence categorization results are shown in Figure 1. A repeated-measures analysis of variance (ANOVA) showed that more faces were categorized as negative when paired with negative information ($F(2, 20) = 6.65$, $p < .01$), whereas more faces were

categorized as positive when paired with positive information ($F(2, 20) = 4.42$, $p < .05$). These results confirmed that participants learned affective value for neutral faces only with one occasion of pairing, contrast to two or four occasions of face-sentence pairs in the original minimal affective learning paradigm [18], and suggested that affective learning is a robust phenomenon.

The liking rating results are shown in Figure 2. A repeated-measures analysis of variance (ANOVA) showed that liking score for faces paired with positive information were higher than for faces paired with positive information and new faces ($F(2, 20) = 6.87$, $p < .01$). This result suggested that the learned positive affective value for neutral faces evoked a more positive affect.

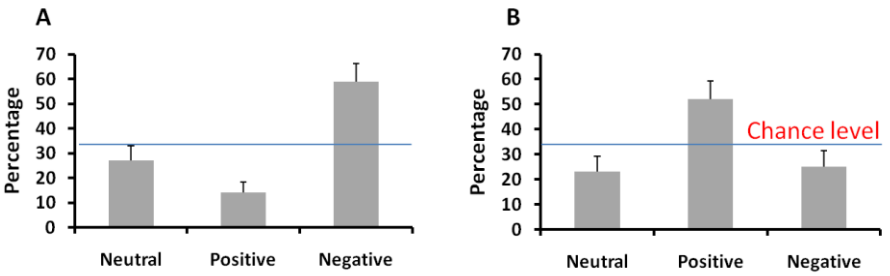


Fig. 1. Affective categorization for faces paired with negative (A) and positive (B) information

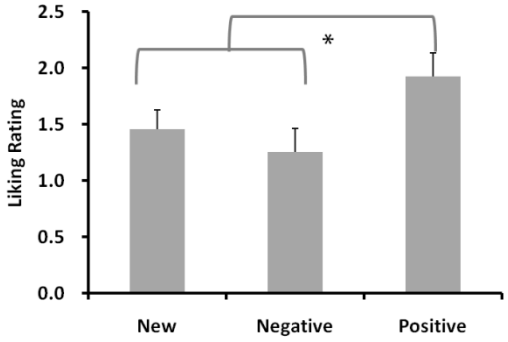


Fig. 2. Affective evaluation scores for new faces and old faces paired with negative and positive information

3.2 The Impact of Affective Learning on Recognition Memory

The old/new judgment data are shown in Figure 3. A repeated-measures analysis of variance (ANOVA) showed that faces paired with positive information are better recognized than with negative information ($F(1, 10) = 8.53$, $p < .05$). This result provided evidence that affective learning do influence the learning performance. Although the accompanied sentence might induce unnecessary processing demands and

distract from learning [12], our result suggested that it is also possible to evoke positive affect, which may, in return, reduce or cancel the additional effort of cognitive activation [19].

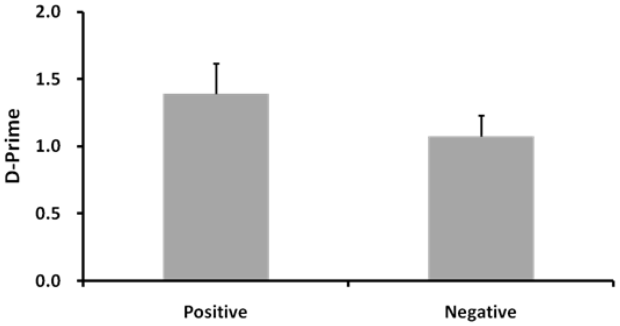


Fig. 3. Recognition memory for faces paired with negative and positive information

3.3 Implications of the Present Study

We found that affective learning occurred for faces paired with affective information, and positive affective learning may improve the learning performance, compared with negative affective learning. Furthermore, only one occasion of pairing can lead to affective learning. This kind of occasion may be very common for the human-computer interactive system. In fact, multimedia learning materials may be designed to associated with affective values (e.g., warm color [10]), and the interface design of learning system can be served as affective companion of the learning materials (e.g., decorative illustrations [11]). In these settings, the learning materials is always paired with affective stimuli, and affective learning is very possible to occur and influence the learning performance. Therefore, affective learning should be taken into account in the interactive system involving human learning.

In human-computer interaction, each main event will be subjectively evaluated by the user, depending on whether the event represents progress or an obstacle towards his or her aims, as appraisal theories imply [20]. The present study provided further evidence that affective context of the event and/or system reactions may also play a highly important role in emotional and learning outcomes. This implication has potential significance to improve several aspects of the interaction such as the interface design, user perception, and task performance. Under affective context, the users may be more stimulated and engaged in the interaction, and can better understand and memorize information because the attention can be modulated by the affective context.

To make machines naturally adapt to their users according to the users' affective state, research on affective computing has focused largely on algorithms that can recognize the affective state of the users. However, Virtual agents that learn naturally from interacting with human may be more essential [21]. According to the finding of this study, affective agents may be served as affective context to modulate the users'

learning. In fact, the use of animated pedagogical agents with emotional capabilities in an interactive learning environment has been found to have a positive impact on learners [22]. Further work may focus on modeling of the affective preferences of the user, the user's emotional response to the interactive interface and agent, as well as the personality and affective responses of the agent to the individual user, and how it will benefit learning performance. It is important that theories and models should concerning affective and cognitive mechanisms used in human computer interaction.

Acknowledgement. This research was supported in part by grants from National Basic Research Program (2011CB302201), and the National Natural Science Foundation of China (31371031).

References

1. Reeves, B., Nass, C.: The media equation: how people treat computers, television,and new media like real people and places. Cambridge University Press, New York (1996)
2. Picard, R.W.: Affective Computing for HCI. In: Proceedings of HCI 1999, pp. 829–833. IOS Press, Amsterdam (1999)
3. Goleman, D.: Emotional intelligence: Why it can matter more than IQ. Random House Digital, Inc. (2006)
4. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
5. Jaques, P., Vicari, R.: A bdi approach to infer students emotions in an intelligent learning environment. Computers & Education 49, 360–384 (2007)
6. Schutz, P.A., Quijada, P.D., de Vries, S., Lynde, M.: Emotion in educational contexts. In: McGaw, B., Peterson, P.L., Baker, E. (eds.) International Encyclopedia of Education, vol. 6, pp. 591–596. Elsevier, Oxford (2010)
7. Ochs, M., Maffiolo, V.: The role of emotions in human-machine interaction. In: Pelachaud, C. (ed.) Emotion-Oriented Systems. ISTE Ltd., London (2012)
8. Cubukcu, F.: The significance of t academic emotions. Procedia - Social and Behavioral Sciences 70, 649–653 (2013)
9. Ben Ammar, M., Neji, M., Alimi, A.M., Gouardères, G.: The Affective Tutoring System. Expert Systems with Applications 37(4), 3013–3023 (2010)
10. Plass, J.L., Heidig, S., Hayward, E.O., Homer, B.D., Um, E.J.: Emotional Design in Multimedia Learning: Effects of Shape and Color on Affect and Learning. Learning and Instruction 29, 128–140 (2014)
11. Magner, U., Schwonke, R., Aleven, V., Popescu, O., Renkl, A.: Triggering situational interest by decorative illustrations both fosters and hinders learning in computer-based learning environments. Learning and Instruction 29, 141–152 (2014)
12. Mayer, R.E.: Cognitive theory of multimedia learning. In: Mayer, R.E. (ed.) The Cambridge Handbook of Multimedia Learning, pp. 31–48. Cambridge University Press, New York (2005)
13. LaBar, K.S., Cabeza, R.: Cognitive neuroscience of emotional memory. Nature Reviews Neuroscience 7, 54–64 (2006)
14. Liu, Y., Fu, Q.F., Fu, X.L.: The interaction between cognition and emotion. Chinese Science Bulletin 54(22), 4102–4116 (2009)

15. Vuilleumier, P., Pourtois, G.: Distributed and interactive brain mechanisms during emotion face perception: Evidence from functional neuroimaging. Neuropsychologia 45, 174–194 (2007)
16. Krolak-Schwerdt, S., Wintermantel, M., Junker, N., Kneer, J.: Reading about persons: The effects of conjunctions on the mental representation of person descriptions. Swiss Journal of Psychology 67(1), 5–18 (2008)
17. Paunonen, S.V.: You are honest, therefore I like you and find you attractive. Journal of Research in Personality 40, 237–249 (2006)
18. Bliss-Moreau, E., Barrett, L.F., Wright, C.I.: Individual differences in learning the affective value of others under minimal conditions. Emotion 8(4), 479–493 (2008)
19. Moreno, R.: Does the modality principle hold for different media? A test of the method-affects-learning hypothesis. Journal of Computer Assisted Learning 22(3), 149–158 (2006)
20. Scherer, K.: Emotion. In: Hewstone, M., Stroebe, W. (eds.) Introduction to Social Psychology: a European Perspective, pp. 151–191. Blackwell Publishers, Oxford (2000)
21. Picard, R.W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Strohecker, C.: Affective learning—a manifesto. BT Technology Journal 22(4), 253–269 (2004)
22. Okonkwo, C., Vassileva, J.: Affective pedagogical agents and user persuasion. In: Stephanidis, C. (ed.) Proceedings of the 9th International Conference on Human-Computer Interaction, pp. 397–401 (2001)

Using Physiological Measures to Evaluate User Experience of Mobile Applications

Lin Yao[1], Yanfang Liu[1], Wen Li[1], Lei Zhou[1], Yan Ge[2], Jing Chai[2,3], and Xianghong Sun[2]

[1] User Experience Lab, China Mobile Research Institute, Beijing, China
{yaolin,liuyanfang,liwen,zhoulei}@chinamobile.com
[2] Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academic of Sciences, Beijing, China
{gey,chaij,sunxh}@psych.ac.cn
[3] University of Chinese Academy of Sciences, Beijing, China

Abstract. Measurements of user experience (UX) in traditional human-computer interaction studies mostly rely on task performance and self-report data. Recent research has showed that physiological measures are good indicators of cognitive involvement and emotional arousal and are suggested being used as a complementary measure of UX. This paper reports a preliminary study to examine the possibility of including physiological measures in the UX evaluation process. In the experiment, participants' physiological responses, task performance and self-report data were collected and analyzed. It was found that physiological measures varied with task performance, as participants showed greater galvanic skin response (GSR) change in the failed tasks than that in the successful tasks. In addition, correlations were found between GSR and self-report data of user experience. The results demonstrated the potential value of physiological measures as a data source of user experience evaluation. However, further investigations involving variations in tasks and individual difference are required.

Keywords: user experience, task performance, self-report, physiological measures.

1 Introduction

Over the past years there has been an increasing interests of user experience (UX) studies for mobile applications. However, trying to measure UX is a task of great challenge as UX is often embedded in a situational, temporal, individual and product context, so it is very difficult to comprehend [1, 2]. Most studies in this field are conducted from the perspective of interface design and ergonomics requirements, using methods like questionnaires, interviews, and heuristic evaluation [3]. In general, these methods are mainly based on two kinds of data: task performance data, such as task completion time and performance error rates, and self-report data of users' personal feelings and preferences [4]. But subjective measures are often reported to be not

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 301–310, 2014.

reliable and required a larger number of subjects and more time for analysis as the scale of an experiment becomes larger [4, 5]. Moreover, both performance data and subjective measures do not directly reflect a users' psychological involvement and fail to explain the cognitive processing and the emotional arousal related. Thus, an objective and more efficient method is needed.

Recently, physiological recordings have been shown to be valuable for measuring cognitive effort and arousal throughout the process of an experience. Physiological measures, such as galvanic skin response, respiration, heart rate, and blood volume pulse, were reported to vary in response to factors such as task difficulty, levels of attention, experiences of frustration and emotionally toned stimuli [6]. Physiological measures as a tool to objectively evaluate user experience have been explored in many studies.

Wilson and Sasse used physiological measures to evaluate subject responses to audio and video degradations in videoconferencing software. Significant physiological responses (increases in GSR and HR, decreases in BVP) were found for videos shown at 5 frames per second versus 25 frames per second even though most subjects didn't notice the difference in media quality, which suggested that physiological measures could be used to uncover the truth which cannot been found from the traditional objective measures of task performance and subjective ratings of user satisfaction [7].

Ward et al. analyzed participants' GSR, BVP and HR data while they attempted to answer questions by navigating through both well and ill designed web pages. It was found that users of the well-designed website tended to relax (indicated by decreases in GSR and HR) after the first minute whereas users of the ill-designed website showed a high level of stress for the most time of the experiment (indicated by increases in GSR and HR) [8].

Mandryk & Inkpen reported studies using psychophysiological recordings to measure user experience with entertainment games. In their experiments, evidence was found that there was a different physiological response when users were playing against a computer versus against a friend. Thee physiological result was also mirrored in the subjective reports provided by participants [9].

In fact, with recent improvements in technology, physiological measures were widely used in other HCI domains, for example, to evaluate presence in stressful virtual environments [10], to measure emotional aspects in mobile contexts [11] and to assess dual-task performance in multimodal human-computer interaction [12].

The goal of this study is to examine the feasibility of relating physiological measures to traditional user experience metrics such as task performance and self-report experience in a mobile context. Tasks are performed on mobile phones to collect data of task performance, self-report user experience and physiological measures. Two research questions are addressed:

- Does physiological measures vary with task performance data when performing tasks on mobile phones?
- Does physiological measures correlate with self-report data of user experience?

2 Method

An experiment was designed to explore whether or not physiological measures are related with traditional user experience metrics. A common used information searching and booking application on mobile phones was selected for the experimental task. In the experiment, scenarios were created to include typical using behaviors, such as finding a restaurant and booking a hotel on the mobile phone. Participants' physiological responses were measured while they are performing the task. Task performance and self-report data of user experience were also collected.

2.1 Experiment Apparatus and Protocol

The experiment was performed in an HCI laboratory. The app was installed on a 4.1-inch Andriod 4.1 smartphone. Physiological data were collected with the BioNeuro Infiniti System and BioGraph Software from Thought Technologies™. GSR and BVP were measured directly by sensors placed on the left fingers. Respiration was measured using a sensor positioned around the thorax. HR was computed from the rawBVP data. All data were collected at 64 HZ. As the BVP sensor is to movement, participants were required not to move their left hand as possible as they can. It should be noted that electroencephalographic (EEG) data and facial expression data were also collected in our experiment and the results were reported in another paper. An experimental scene was show in Figure 1.

Fig. 1. An experimental scene: physiological measures were collected when the participant performing tasks on a mobile phone

Upon arriving, participants signed a consent form, after which they were fitted with the physiological sensors and were allowed to have free use of the mobile phone for approximately three minutes. Then, a three-minute resting baseline for physiological measures was gathered before the experimental tasks. There were five tasks in total. After each task, participants rated the level of task difficulty on a 5-point scale, and

after finishing all the five tasks, they were asked to answer a questionnaire of overall user experience upon the application. The whole procedure of the experiment took about 30 minutes on average.

The five tasks were to book a hotel (task A), to check for the location of a given restaurant (task B), to find out the best route to the restaurant (task C), to make a comment on the restaurant (task D) and to search for a KTV in the nearby (task E). Prior to the formal experiment, a pilot-study was carried to ensure that these tasks were of appropriate difficult.

2.2 Participants

Ten males and ten females were recruited to participate in the experiment. Most of them have little or no experience with the app. Physiological data was missing for one female participant, so data from nineteen participants aged 28-35 was analyzed and presented.

2.3 Measurements

Task Performance. Task performance was measured by task completion rate, that is, the proportion of participants who successfully completed the task. Besides, perceived task difficulty was also assessed by scores on five-point Likert scale (1= not difficult at all, 5 = very difficult).

Self-report of User Experience. The overall user experience of the mobile application was assessed with the User Experience Questionnaire (UEQ) [13]. The UEQ was developed as a tool for the quick assessment of the user experience of interactive products. It consists of 26 bipolar items which to be rated on a seven-point Likert scale. The 26 items are assigned to six dimensions: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. The final score of each of the six dimension was scaled from -3 to +3.

Physiological Measures. Based on previous research, physiological parameters assessed in this study were galvanic skin response (GSR), blood volume pulse (BVP), heart rate (HR) and respiration rate, which were proved to be good indicators of stress and arousal (See [14] for a review).

GSR is a measure of the skin conductance. It varies linearly with the overall level of arousal and increases with anxiety and stress and is considered as a reliable indicator of affective response [15].

BVP signal is an indicator of blood flow. It increases with negative valence emotions such as fear and anxiety, and decreases with relaxation [16].

HR in another measure of cardiovascular activity which reflects emotional state. It has been found to increase for a number of negative emotions (e.g. anger, anxiety, embarrassment, fear, crying sadness) as well as for some positive emotions (e.g. happiness, joy) and surprise [17].

Respiration is measured as the rate of volume at which an individual exchanges air in their lungs. Previous research has found that emotional arousal increase respiration rate while rest and relaxation decreases respiration rate [18].

Since there was a large individual difference in physiological signals, individual baseline have to be taken into account. As adopted by similar studies, normalized GSR, BVP, HR and respiration rate on each task were calculated using the formula (signal-baseline)/baseline for each participant [4, 8, 9].

3 Results

The result section consisted of two parts. First, normalized physiological data in different tasks was presented and analyzed with the task performance data. Then, correlations between overall physiological response and self-report user experience were examined. All the data were submitted to SPSS 20.0 for analysis.

3.1 Task Performance and Physiological Data

Means of task completion rates and perceived task difficulty were shown in Figure 2 and Figure 3, respectively. A repeated measures ANOVA was used to analyze the data. The results showed significant differences among different tasks, F (4, 76) = 13.36, $p < .001$. Multiple comparisons showed that task completion rates of task B and task E were significantly higher then the other three tasks, $ps < .001$. The results of perceived task difficulty were in line with task completion rates, that is, tasks with lower completion rates, such as task A, C and D, were reported to be more difficult.

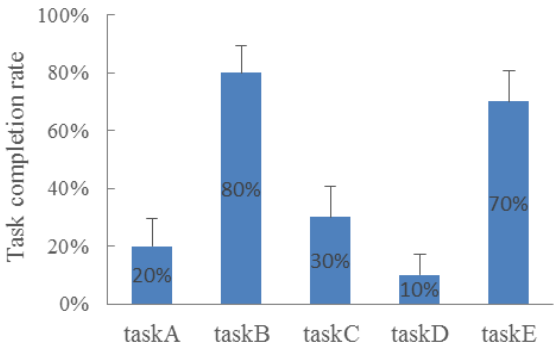


Fig. 2. Task completion rates of the five tasks. Error bar stands for one standard error.

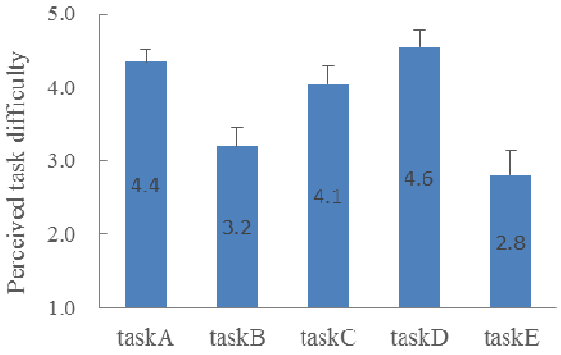


Fig. 3. Perceived task difficulty of the five tasks (average scores on five-point Likert scale, 1= not difficult at all, 5 = very difficult). Error bar stands for one standard error.

The physiological data were also analyzed across tasks. The results showed that tasks with lower task completion rates (which were also perceived as more difficult) tended to cause greater normalized GSR (see Figure 4), but the trend failed to be significant, $F(4, 76) = 1.28$, $p > .05$ (repeated measures ANOVA). The lack of statistical significance might contributed to two reasons. First, the number of participants was not large enough to distinguish subtle GSR differences between different tasks. Second, participants were not consistent on which task was more difficult or easier, thus, the effect of difference caused by task difficulty might be reduced when averaged across individuals.

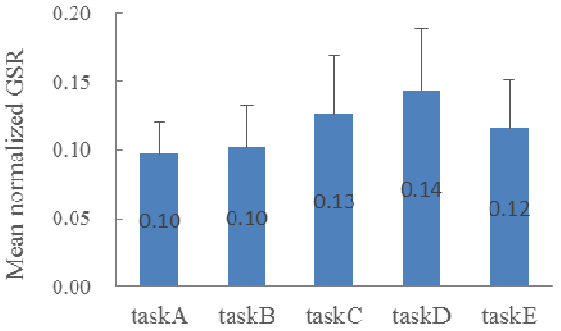


Fig. 4. Mean normalized GSR of the five tasks. Error bar stands for one standard error.

To further investigate the relationship between task performance and physiological data, the mean normalized GSR in the successful tasks was calculated and compared to that in the failed tasks for each participant. Means of normalized GSR across the two types of tasks were shown in Figure 5. The repeated measures ANOVA analysis showed a marginal significant difference on mean normalized GSR between successful tasks and failed tasks, $F(1, 14) = 4.17$, $p = .061$ (repeated measures ANOVA,

only 14 pairs of normalized GSR data were obtained as four participants succeed in all of the five tasks and had no data for failed tasks). Participants showed greater normalized GSR when they failed the task, which was in consist with previous research that GSR data was sensitive to the stress caused by the task difficult [4].

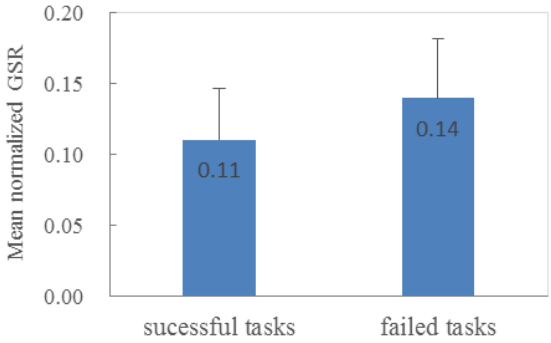


Fig. 5. Mean normalized GSR in successful tasks and failed tasks. Error bar stands for one standard error.

The BVP, HR, and respiration rate data were also analyzed in a similar way but no significant difference was found.

3.2 Self-report User Experience and Physiological Data

The six-dimension UEQ scores when using the mobile application were shown in Table 1. The overall user experience was somehow negative as all six-dimension scores were below zero.

Table 1. The overall user experience indicated by six-dimension UEQ scores (data were presented as means of score scaled from -3 to +3)

UEQ Score	Mean	SD
Attractiveness	-1.30	1.11
Perspicuity	-1.16	1.18
Efficiency	-1.25	1.24
Dependability	-0.92	0.97
Stimulation	-0.74	0.90
Novelty	-1.07	1.15

Correlation analysis showed that attractiveness, efficiency, dependability and novelty were significantly correlated with GSR (rs ranged from 0.46 to 0.58, all ps < .05, see Figure 6), which indicated that physiological measurement such as GSR, to some extent, revealed subjectively reported user experience.

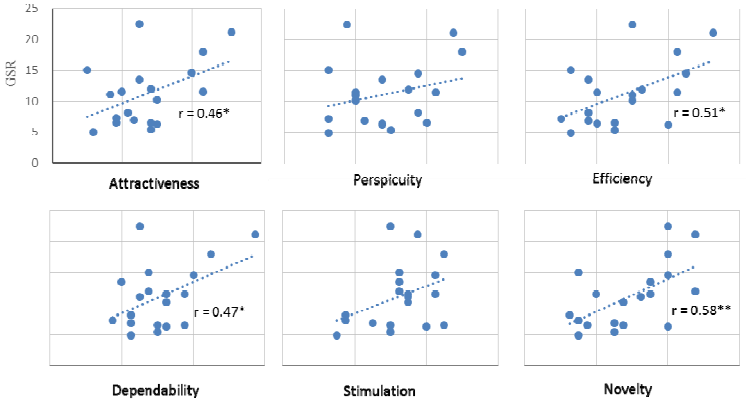


Fig. 6. Correlations between UEQ scores and GSR

4 Discussions, Conclusions, and Future Work

In this study, participants' physiological response when using a mobile application were collected and analyzed with both task performance data and self-report data of user experience. There were two major findings. The first was that physiological measures vary with task performance as participants showed greater GSR change in the failed tasks than that in the successful tasks. Second, significant correlations was found between GSR and subjective assessment of user experience (such as attractiveness, efficiency, dependability and novelty).

Lin and Hu suggested to establish a new UX evaluation method based on three kinds of data: task performance, subjective assessment data and physiological data [4]. Though our results demonstrate the possibility of correlating physiological data with the other two types of data, further investigations are needed.

First, our experiment showed that physiological data (GSR) correlated to task performance and subjective assessment, more rigorous experimental control and analytical methods are need to understand how these relationships are established. For example, in future studies, physiological response may be synchronized with behavior observation method such as eye-tracking technology to examine the relationship between behavior response and users' current psychophysiological state, through which we are able to understand what problems are and how user react to them.

Second, one of the most unique feature of mobile applications is that they are typically used in a changing context. Results from laboratory studies are questioned as users' experience of interaction with products varies greatly with the context and the sensitive of physiological measures decreases in movement context. Therefore, experiments should be extended to more ecologically valid context and a variety of tasks.

Third, a large individual difference exists in physiological data and studies including more participants are needed to ensure the power of statistical tests.

In sum, our study found that participants' physiological responses correlated with task performance and self-report data when they were using a mobile application. The results, though being preliminary and requiring further investigation, suggest the potential value of physiological data as a data resource for user experience evaluation.

Acknowledgement . This work was supported by User Experience Lab of China Mobile Research Institute, NSF China (31100750, 91124003) and Science and Technology (S&T) basic work (2009FY110100).

References

1. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.-B.: User experience over time: an initial framework. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 729–738. ACM (2009)
2. Zhang, D., Adipat, B.: Challenges, methodologies, and issues in the usability testing of mobile applications. International Journal of Human-Computer Interaction 18, 293–308 (2005)
3. Tullis, T., Albert, W.: Measuring the user experience: collecting, analyzing, and presenting usability metrics. Morgan Kaufmann, San Francisco (2010)
4. Lin, T., Omata, M., Hu, W., Imamiya, A.: Do physiological data relate to traditional usability indexes? In: Proceedings of the 17th Australia conference on Computer-Human Interaction, pp. 1–10. ACM (2005)
5. Annett, J.: Subjective rating scales: science or art? Ergonomics 45, 966–987 (2002)
6. Andreassi, J.L.: Psychophysiology: Human behavior and physiological response. Psychology Press, Kentucky (2000)
7. Wilson, G.M.: Psychophysiological indicators of the impact of media quality on users. In: CHI 2001 Extended Abstracts on Human Factors in Computing Systems, pp. 95–96. ACM (2001)
8. Ward, R.D., Marsden, P.H.: Physiological responses to different WEB page designs. International Journal of Human-Computer Studies 59, 199–212 (2003)
9. Mandryk, R.L., Inkpen, K.M., Calvert, T.W.: Using psychophysiological techniques to measure user experience with entertainment technologies. Behaviour & Information Technology 25, 141–158 (2006)
10. Meehan, M., Insko, B., Whitton, M., Brooks Jr., F.P.: Physiological measures of presence in stressful virtual environments. ACM Transactions on Graphics (TOG), 645–652 (2002)
11. Ganglbauer, E., Schrammel, J., Geven, A., Tscheligi, M.: Possibilities of Psychophysiological Methods for Measuring Emotional Aspects in Mobile Contexts. In: MobileHCI 2009, p. 15. ACM (2009)
12. Novak, D., Mihelj, M., Munih, M.: Dual-task performance in multimodal human-computer interaction: a psychophysiological perspective. Multimedia Tools and Applications 56, 553–567 (2012)
13. Laugwitz, B., Held, T., Schrepp, M.: Construction and Evaluation of a User Experience Questionnaire. In: Holzinger, A. (ed.) USAB 2008. LNCS, vol. 5298, pp. 63–76. Springer, Heidelberg (2008)

14. Forne, M.: Physiology as a Tool for UX and Usability Testing. School of Computer Science and Communication, Master. Royal Institute of Technology, Stockholm (2012)
15. Hudlicka, E.: Affective Computing: Theory, methods, and applications. CRC Press, Boca Raton (2011)
16. Healey, J.A.: Wearable and automotive systems for affect recognition from physiology. Massachusetts Institute of Technology (2000)
17. Kreibig, S.D.: Autonomic nervous system activity in emotion: A review. Biological Psychology 84, 394–421 (2010)
18. Stern, R.M., Ray, W.J., Quigley, K.S.: Psychophysiological recording. Oxford University Press, New York (2001)

Cognitive Psychology in Aviation and Space

Applying Cognitive Work Analysis to a Synthetic Aperture Radar System

Kerstan Cole, Susan Stevens-Adams, Laura McNamara, and John Ganter

Sandia National Laboratories
P.O. Box 5800, Albuquerque, NM 87185, USA
{kscole,smsteve,lamcnam,jganter}@sandia.gov

Abstract. The purpose of the current study was to analyze the work of imagery analysts associated with Sagebrush, a Synthetic Aperture Radar (SAR) imaging system, using an adapted version of cognitive work analysis (CWA). This was achieved by conducting a work domain analysis (WDA) for the system under consideration. Another purpose of this study was to describe how we adapted the WDA framework to include a *sequential* component and a means to explicitly represent relationships between components. Lastly, we present a *simplified work domain representation* that we have found effective in communicating the importance of analysts' adaptive strategies to inform the research strategies of computational science researchers who want to develop useful algorithms, but who have little or no familiarity with sensor data analysis work.

Keywords: Cognitive Work Analysis, Work Domain Analysis, Human Factors, Synthetic Aperture Radar, Imagery, Systems Analysis.

1 Introduction

Remote sensing domains are common and complex work domains comprising multiple subsystems, components and actors. Such systems provide society with a wide range of information products, from space weather to patterns of change in land use. As remote sensing platforms become more sophisticated, the human actors responsible for managing and analyzing data feeds are increasingly facing a "data deluge" that will inevitably change how data consumers interact with the information products derived from remotely sensed systems. Automated support for analysis is a necessary evolution. However, because sensor data analysis is a highly interpretive process shaped by contextually-specific goals, automated analytical systems present significant design challenges for algorithm and software developers.

In this paper, we discuss the use of cognitive work analysis (CWA) methods, specifically work domain analysis, and the construction of an abstraction hierarchy, to decompose one sub-domain of a remote sensing data analysis workflow associated with the Sagebrush system, a Synthetic Aperture Radar (SAR) platform used to generate ground image data for a wide range of civilian and military applications. We provide a brief overview of the Sagebrush system and summarize the research and

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 313–324, 2014.

design challenges that motivated our study of one sub-domain in the larger Sagebrush workflow. We then describe some of the difficulties we encountered in using CWA representations to communicate our findings. This difficulty motivated us to evolve some elements of CWA into representations and terminology that enhance understanding of the methodology's power. We discuss how our products are being used by algorithm and software experts to develop and evaluate new algorithms, software and visualization platforms to enhance the analysis of SAR data and image products.

2 The Sagebrush System

Sagebrush refers to a family of SAR sensors that are used to support a wide array of civilian and military ground operations in multiple locations throughout the world. Taken in its entirety, Sagebrush is a large, complicated work domain that includes a wide array of sites, operators and analysts, platforms, networks, locations, workstations, offline and off-site databases, communications platforms, qualitative and quantitative data, and copious amounts of imagery.

The research that we are pursuing focuses on the perceptual and cognitive work of imagery analysts associated with Sagebrush sensing platforms. As is true with most remote sensing systems, data generated by the Sagebrush system goes through several stages of processing, review, information extraction, and knowledge product creation. Analysts at the front-end of the sensor perform – the so-called "near-real time" analysts – perform rapid triage, assessment and communication of trends and events and trendss for Sagebrush's stakeholder community. Other groups of analysts work with Sagebrush products in an "offline" process that generates longer-term, strategic assessments of trends and events. Such offline analyses shape the planning and implementation of Sagebrush missions and even the development of next-generation Sagebrush hardware and software.

The CWA activities describe in this paper focused on the domain of "offline analysis." Offline Sagebrush analysts are responsible for assessing the correctness, completeness and overall performance of fielded Sagebrush systems. Their job involves not only analysis of Sagebrush data products, but also the incorporation of several types of auxiliary data (e.g., weather, agricultural activity, animal movement) to develop richer evaluations of trends and events rendered in the sensor data. Associated tasks include the retrospective analysis of radar imagery data (i.e., analyzing the features of an image that contains evidence of ground changes or signatures of ongoing trend); classification of events and trends captured in the imagery; evaluation of the periodicity of scene changes to identify emerging trends; and helping fielded Sagebrush teams improve their performance with richer contextual data for trend analysis. The specific tasks associated with this offline analytic workflow are labor intensive, cognitively demanding, and require extensive domain knowledge about the sensor, the terrain being imaged, the operational requirements of fielded teams, and the needs and requirements of Sagebrush stakeholder groups.

Figure 1 shows the basic flow of information and data from the sensor to the offline analytic domain described in this paper. As shown, the Sagebrush sensor gene-

rates radar data that is rendered in pixelated imagery. This imagery is then transmitted to both off-site and local servers for storage. In addition, analytic teams deployed with these platforms perform "near-real time" triage of data products identify and characterize events and trends for Sagebrush stakeholders. These near-real time products and associated imagery and sensor data are then transmitted to a variety of other consumers, including Sagebrush offline analysts. Together, these sensor data, imagery and near-real time analytic products comprise the critical information resources for Sagebrush's offline analytic work.

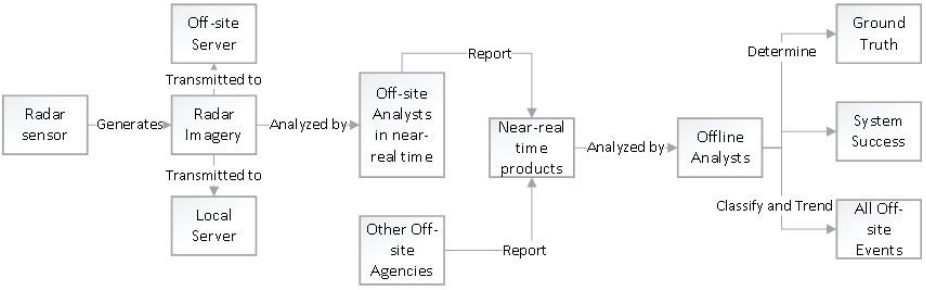


Fig. 1. Basic Work Flow for Sagebrush Data and Imagery Analysis

Every day, Sagebrush offline analysts log into their computers to determine if new image products, reports, and sensor data are waiting for review and evaluation. They also review all events and trends reported by other off-site agencies. This requires a review of radar imagery generated by their sensors and reports generated by off-site agencies. One result of this analysis is a list of all the events and trends that have occurred during a particular time period in an area being imaged by Sagebrush radar platforms, creating a set of assessments that the analysts describe as 'ground truth.' Sagebrush owners use this ground truth to determine the effectiveness of their system in meeting all stakeholder information requirements. In addition, the offline analysts continually revise and update a database of ground trends and events that can be disseminated to a much wider operational community. Essentially, Sagebrush's offline analysts take products generated by their counterparts and put them into a broader, longer-term context that enables not only evaluation of fielded system performance, but the enrichment of the entire Sagebrush community's collective knowledge about operationally-significant events and trends in the areas under study.

2.1 Motivating Context for This Research

Because SAR systems provide all-weather sensing capabilities and are relatively easy to mount on a variety of airborne platforms, they are becoming increasingly popular for a wide range of remote sensing tasks. As the volume and diversity of SAR imaging missions expands, stakeholders are grappling with floods of sensor data and are seeking new ways to analyze sensor data, beyond the standard "eyes-on-imagery" paradigm that dominates remote sensing analysis. Our team is part of a larger project called **PANTHER** – **P**attern **AN**alytics to support **H**igh-performance **E**xploitation and **R**easoning – funded by Sandia National Laboratories in Albuquerque, NM.

PANTHER researchers are pursuing new algorithms, software architectures, and visualization platforms to enable human analysts to realize the information value of remotely sensed datasets. Studies of working analysts are critical to understanding how humans interact with sensor datasets, so that software designers can develop usable, useful and adoptable technologies that demonstrably enable people to extract meaningful information from these datasets.

3 Cognitive Work Analysis

Within PANTHER, our team was tasked with studying the current work processes of Sagebrush analysts and generating ideas and requirements for algorithms, architectures and visualizations to enhance analytic work. To address this challenge, we conducted a CWA study. CWA is an evolution of cognitive task analysis (CTA) methods that was specifically designed for complex systems with uncontrolled, uncertain environments [1-3]. CTA provides detailed analysis of discrete, predefined task sequences performed by individuals; in contrast, CWA decomposes an entire work domain, then asks questions about how operators navigate toward domain-specific goals using resources at hand. In doing so, CWA reveals the creative work of domain experts operating complex systems under conditions of uncertainty and constraint, and how we can design systems in ways that will enhance operator performance. CWA is increasingly recognized as a valuable framework for eliciting and documenting the human activities associated with a technological system: the tasks and activities that human operators perform, the behavior resulting from their interaction with the system, their work context, and the goals and purpose that motivate their actions [2], [3].

Additionally, although CWA has been applied to many domains, a recent review [4] indicates that sensor data analysis – a highly visual and individualized form of work – is not one of them. Thus, one purpose of the current study was to evaluate the usefulness of CWA approaches, specifically work domain analysis, for informing the design of statistical and graph-based algorithms to mine patterns in very large sensor datasets. A second purpose was to describe how we adapted the work domain analysis framework, as proposed by Vicente [3] and Naikar et al. [2], to include a *sequential* component, a means to explicitly represent relationships between components, detailed explanations of the different abstractions that exist within a system hierarchy, and how outputs from this analysis can be used as direct inputs for a system interface. Lastly, we present a *simplified work domain representation* that we have found effective in communicating the importance of analysts' adaptive strategies to inform the research strategies of computational science researchers who want to develop useful algorithms, but who have little or no familiarity with sensor data analysis work.

3.1 Work Domain Analysis (WDA)

Overview. Lintern describes a work domain as "an intentional-functional-physical space in which work can be accomplished."[5] He explains that intention refers to the system's purpose and that function denotes an "activity-independent capability to accomplish something specific." Essentially, WDA is a means for practitioners to

identify the purposes and constraints of a system and to describe system components, and their interactions and relationships in operators' work. WDA is the first phase of CWA and has been used in a variety of domains to inform system interface design (for a review, see [4]). The representational product of WDA is an abstraction hierarchy (AH). This tool is a hierarchical representation that describes the system in terms of its functional purpose, values and priority measures, purpose-related functions, object-related processes and physical objects. The following is a summary of the different levels of abstraction proposed by [2]:

Table 1. The abstraction axis of the Abstraction Hierarchy

Abstraction Level	Description
Functional Purpose	The purposes of the work system and the external constraints on its operation
Values and Priority Measures	The criteria that the work system uses for measuring its progress towards the functional purpose
Purpose-related Functions	The general functions of the work system that are necessary for achieving the functional purpose
Object-related Processes	The functional capabilities and limitations of physical objects in the work system that enable the purpose-related functions
Physical Objects	The physical objects in the work system that afford the object related processes

4 Completing the WDA

We adapted the nine steps proposed by [2] for completing a WDA. The abstraction hierarchy was developed as a tool to deconstruct the work domain.

1. **Establish the purpose of the analysis:** The purpose of this analysis was to deconstruct the offline synthetic aperture radar work domain in order to determine if operator goals and tasks are currently supported by the system and to develop design recommendations for tools that support these goals.
2. **Identify the project constraints:** The project was constrained by the type of analysis tools that the authors could use to observe offline radar imagery analysts' work. Analysts work in a classified environment. Thus recording software tools were prohibited. Other project constraints included resource constraints and time.
3. **Identify the boundaries of the analysis:** This analysis focuses solely on the work domain in which offline analysts perform.
4. **Identify the nature of the constraints in the work domain:** The timeframe in which analysts perform their duties varies. Sometimes, analysts are unable to perform their work because imagery is absent or equipment is especially slow. Procedural work constraints exacerbate this problem. It may take days or weeks for technicians to fix software and/or hardware related issues. Other constraints include political constraints, and mission constraints which are outside the scope of

this analysis. Analysts have a specific set of operationally defined work requirements. These are explicit and analysts do not deviate from them.

5. **Identify the sources of information for the analysis:** We interviewed and observed two offline radar imagery analysts for approximately 50 hours. Analysts performed verbal walkthroughs of their work for many different types of imagery events and trends. These walkthrough also included discussions about imagery, different reports, and analyst-developed software applications. We also attended analysts' weekly meetings where they are briefed about factors that may influence the way that they perform their work
6. **Construct the AH with readily available sources of information:** The AH was constructed with the sources of information described in step 5.
7. **Construct the AH by conducting special data collection exercises:** The data collection exercises included structured interviews, observations, and attendance at analysts' weekly meetings. Data collection lasted over a period of months from April until August of 2013.
8. **Review the AH with domain experts:** We asked analysts to provide feedback about the accuracy of our representations of their tasks, work flow etc. throughout the entire study.
9. **Validate the AH:** We plan to complete this step during a future study.

5 Results

Figure 2 shows a completed AH for the system under consideration based on [2] and [3].

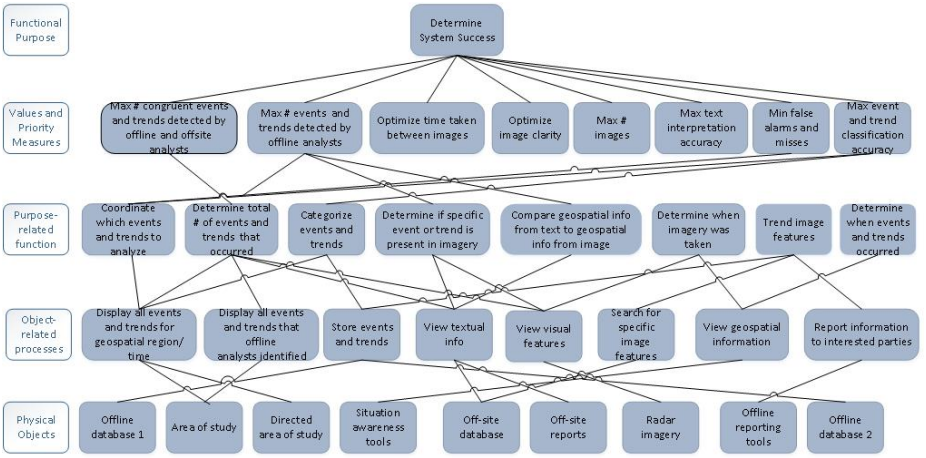


Fig. 2. Abstraction Hierarchy for offline analysis of the Sagebrush SAR system

We began constructing this representation by examining the physical objects that analysts use (Figure 3). This layer is shown at the bottom of the hierarchy and in-

cludes software objects such as databases, scripts, and offline applications. One-to-one mappings between tools and different analyst processes do not exist within this system. Instead, each process uses a selection of tools that overlaps with other processes.

Fig. 3. Physical objects of the system

We then spoke to analysts about how these objects are used to accomplish specific processes (Figure 4). For example, the offline database serves as storage for events and trends and a means for offline analysts to track the progress of their analysis. In addition, the area of study for all events and trends and the directed area of study for events and trends are represented by visualizations that allow analysts to obtain a cursory understanding of time and place for imagery features.

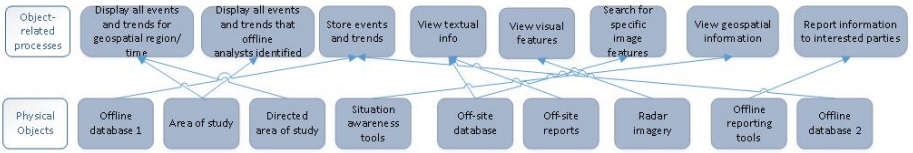


Fig. 4. The physical objects and associated processes layer of the AH

The purpose-related function layer of the abstraction hierarchy (Figure 5) consists of higher-level functions that are associated with object-related processes and the values and priority measures of the system. For example, analysts view geospatial information of image features in order to determine when the imagery was taken. They view the visual features in order to determine if specific events and trends are present in the imagery. This layer often reveals gaps between system functions and their associated object-related processes and priority measures.

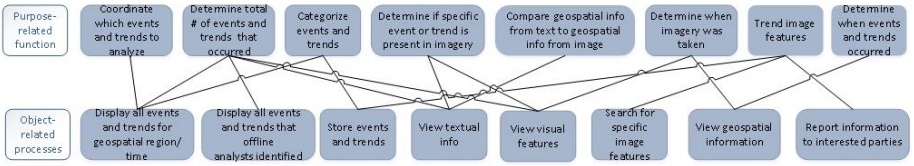


Fig. 5. Object-related processes and their purpose-related functions

Figure 6 shows how the purpose-related functions of the system can be accomplished through a set of values and priority measures. The values and priority measures of the system have great utility in terms of characterizing and sometimes measuring behaviors in complex systems. As shown, one way to achieve system success is to minimize the false alarm and miss rate for offline analysts. Analysts achieve

this through correctly determining when events and trends occurred, by determining when the imagery was taken, and by coordinating events among other analysts.

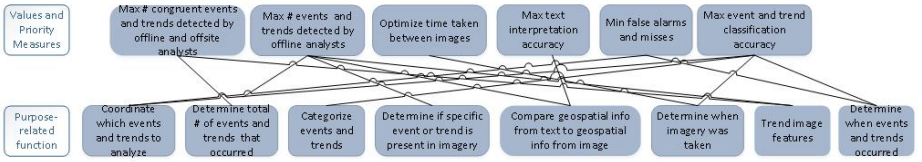


Fig. 6. The values and priority measures of the system's purpose-related functions

The top of the AH shows the functional purpose of the system. Although there are several purposes, we have only reported one due to space limitations. As shown in Figure 7, the functional purpose of the system is to determine system success. This is measured by all of the values and priority measures shown.

Fig. 7. The functional purpose of the AH is measured through the values and priority measures

5.1 Evaluation of WDA Framework

Although [2] and [3] provide a framework that has great utility for representing complex systems, it is not a universal solution for every domain. By showing the connections between the different layers, one can certainly see how system components are related on a higher level. However, one cannot make a determination about the quality of these relationships nor can they ascertain whether the system adequately supports particular functions and processes. Similarly, most WDA practitioners do not include contextual features such as sequence. Sequential steps are usually analyzed independently of WDA analysis during CTA or HTA. However, these methods can complement WDA. Sequence may provide more context for design requirements.

Moreover, although the nine-step methodology developed by [2] provides an overview of the steps required to perform WDA, the steps are ambiguous at best. The particular details of the steps are lacking. For example, step 6 states to complete the AH with readily available information. However, it gives no further guidance about how to do this. New practitioners would be unlikely to know where to begin.

5.2 Adapted WDA Framework

In order to accomplish our analysis of the system under consideration, we added details to the original WDA methodology for completing an AH. Firstly, as mentioned previously, step 6 states to complete the AH with readily available information. We

suggest beginning this step by populating the bottom of the AH. This can be accomplished by creating an inventory of the system's physical objects and their associated processes. Afterwards, complete the top of the AH by determining the functional purpose of the system. The middle layers are easily the most difficult to understand and represent. Further guidance is needed to move these levels from a philosophical framework to more concrete representations that can inform design.

We also suggest the addition of a step between 7 and 8 in Naikar's nine-step methodology: construct complimentary data representations. After constructing the AH based on information obtained from the data collection exercises, we organized the components of the hierarchy by the sequence in which they are used and by their function rather than through a vertical dimension proposed by [3]. Then, we completed the bottom of the AH by conducting a separate hierarchical task analysis [6]. We grouped physical objects by function and by sequence to provide more context to develop system design requirements. In addition, we represented the relationships between items in the AH (e.g., not supported, weakly supported, adequately supported) by drawing different types of lines (e.g., dashed lines represent weak support, solid lines denote adequate support, and missing lines denote a lack of support). This also allows a level of system transparency that is not present in previous frameworks. Essentially, it allows practitioners to easily see the relationships between components of the work domain without extensive interpretation.

Figure 8 is an AH completed using the adapted framework suggested above:

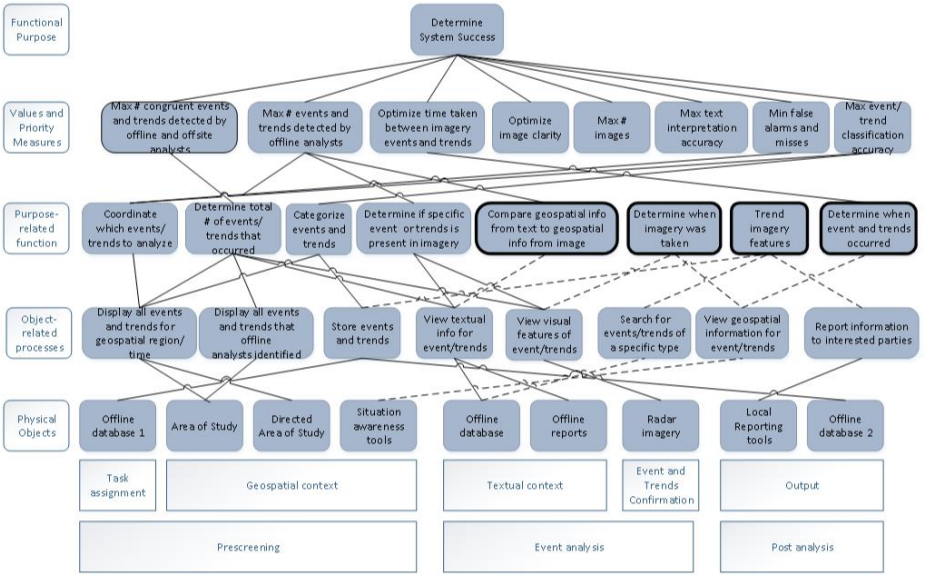


Fig. 8. Adapted HA for a synthetic aperture radar system

As shown in Figure 8, we created an inventory of the physical objects of the system. This is represented as the bottom layer of the hierarchy. These are mostly software objects. We ordered this inventory to correspond to the operating sequence of

the system: prescreening, events and trends analysis, and post-screening. Objects were also grouped according to function including task assignment, geospatial context, textual context, events and trends confirmation, and outputs.

As mentioned previously, the lines between the layers represent relationships between objects, processes, functions, measures and purposes. Traditional CWA represents all of these relationships by drawing the same solid line. However, all relationships are not equal. Thus, by depicting the differences between these interactions, practitioners can more easily determine areas for improvement. For example, offline database 1 is used to store events and trends. Essentially, it is a spreadsheet that contains a list of events and trends that have occurred within a particular geospatial timeframe. The solid line indicates that this function (store) is adequately supported by its tool (database). However, searching for specific events and trends types is weakly supported by the tools in the current system. Offline analysts can search only for a subset of events and trends types, which excludes many other types. A dashed line represents the weak support for this process. Similarly, the situation awareness tools require manual transfer of information and von-screen visually matching. This may increase the likelihood for human error. Thus, this relationship was designated as weak because the system does not optimize human capabilities and limitations for this process. Thus, the line is dashed between the off-site reports and searching for events and trends types.

The bolded boxes show functions that are weakly supported by the current system. This is perhaps where the most improvement can occur. For example, one function of the system is to trend image features. However, analysts' ability to do this is inhibited by the tools they use and the processes that allow them to complete their work. Although they may be able to trend particular features, the system does not represent or catalog the full suite of imagery features.

5.3 A simplified Explanation of the AH

Unless one has extensive experience creating and reading abstraction hierarchies, it is often difficult to understand the messages they convey. We suggest a simplified explanation of the AH developed by Ganter [7]: As shown in Figure 9, the AH shows the connections between why the system exists at the top (i.e., its functional purpose), what it consists of (i.e., values and priority measures, purpose-related functions), and how it functions (e.g., physical objects) at the bottom. Essentially, it is a system hierarchy. Each level of the hierarchy has a different time horizon [7]. The functional purpose of a system evolves slowly often over years. This includes the mission, goals and constraints of the system. Similarly, the physical objects at the bottom of the hierarchy also evolve slowly because this change requires both design and execution of this design.

However, the middle of the hierarchy can evolve quickly. We adopt Ganter and colleagues' definition of the collection of middle phases as the zone of adaptation (see Figure 9). It is in this zone of adaption where operators can enact change quickly by adjusting goals and tasks [7]. In effect, the human actors adjust and revise their mental models of the system through dialog and learning. By examining this zone of

adaptation, we can learn what operators do with new system capabilities to achieve enduring goals. These changes may in turn suggest new ways to levy engineering capabilities.

Table 2. The Abstraction Hierarchy decomposes a system into why, what and how layers with different times scales [7]

Why	Functional Purpose Mission, goals, constraints	Evolves slowly (years)
What	Zone of Adaptation: object related processes, purpose-related functions, values and priority measures Goals: what needs to be accomplished Tasks: Actions by operators to achieve goals	Changes rapidly in response to situations and events
How	Physical Objects Hardware, software, algorithms	Evolves slowly (days to months)

Acknowledgments. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

1. Jenkins, D., Stanton, N., Walker, G., Salmon, P., Young, M.: Creating interoperability between the Hierarchical Task Analysis and the Cognitive Work Analysis Tool. Report from the Human Factors Integration Defence Technology Centre, U.K. (2006)
2. Naikar, N., Hopcraft, R., Moylan, A.: Work domain analysis: Theoretical concepts and methodology: Australian Defence Science and Technology Organisation, Report DSTO-TR-1665 (2005)
3. Vicente, K.J.: Cognitive Work Analysis: Toward Safe, Productive, and Health Computer Based Work. Lawrence Erlbaum Associates Inc., Mahwah (1999)
4. Read, G.J.M., Salmon, P.M., Lenne, M.G.: From work analysis to work design; A review of cognitive work analysis design applications. In: Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting, pp. 368–371 (2012)
5. Lintern, G.: Work-focused analysis and design. Cognition, Technology, & Work 14, 71–81 (2012)
6. Stevens-Adams, S., Cole, K., McNamara, L.: Hierarchical task analysis of a synthetic aperture radar process. In: The Proceedings of the Human Computer Interaction International Conference (2014)

7. Ganter, J.H.: Cognitive Systems Engineering: Work Domain Analysis for an Evolving Space Sensor Operation. SAND Report 2013-9017P released on Intelink (2013)
8. Burns, C.M., Bryant, D.J., Chalmers, B.A.: Boundary, purpose and values in work-domain models: Models of naval command and control. IEEE Transactions on Systems, Man and Cybernetics 35, 603–616 (2005)
9. Miller, A., Scheinkestel, C., Steele, C.: The effects of clinical information presentation on physicians' and nurses' decision-making in ICUs. Applied Ergonomics 40, 753–761 (2009)
10. Salmon, P.M., Regan, M., Lenne, M., Stanton, N.A., Young, K.: Work domain analysis and intelligent transport systems: Implications for vehicle design. International Journal of Vehicle Design 45, 426–448 (2007)

The Investigation of Pilots' Eye Scan Patterns on the Flight Deck during an Air-to-Surface Task

Wen-Chin Li[1,*], Graham Braithwaite[1], and Chung-San Yu[2]

[1] Safety and Accident Investigation Centre, Cranfield University, Martell House, University Way, Cranfield, Bedfordshire, MK43 0TR, United Kingdom
[2] Department of Industrial Engineering and Engineering Management, National Tsing Hua University, R.O.C.
wenchin.li@cranfield.ac.uk

Abstract. Twenty qualified mission-ready F-16 pilots participated in this research. The ages of participants are between 26 and 46 years old (M=33, SD=6); total flying hours between 400 and 3,250 hours (M=1358, SD=882); F-16 type flying hours between 101 and 2,270 hours (M=934, SD=689). Eye movement data were collected by a head-mounted ASL (Applied Science Laboratory) Mobile Eye which was 76 grams in weight, combined with F-16 flight simulator, a dynamic high fidelity trainer that replicates actual aircraft performance, navigation and weapon systems. The scenario is an air-to-surface task. Participants have to intercept the proper route and turn toward the target at an altitude of 500 feet with speed of 500-KIAS, then performing a steep pop-up manoeuver to increase altitude abruptly for appropriate reconnaissance, following by dive and roll-in toward the target to avoid hostile radar lock-on. When approaching the target, subjects have to roll-out, level the aircraft, aiming at the target, release the weapon, and finally pull-up with a 5～5.5 G-force to break-away from the range. The results show significant differences in pilots' number of gaze points among five different AOIs, $F(4, 95) = 533.84$, $p<.001$, $\eta 2\rho = .97$. Further comparisons using post-hoc Bonferroni adjusted tests showed HUD has a significantly higher numbers of gaze points than ICP, DED, RMFD and LMFD; and ICP has significantly higher gaze points than DED, RMFD and LMFD. Also, there were significant differences in pilots' number of fixation among five different AOIs, $F(4, 95) = 306.98$, $p<.001$, $\eta 2\rho = .94$. Further comparisons using post-hoc Bonferroni adjusted tests showed HUD has significantly higher number of fixation than ICP, DED, RMFD and LMFD; and ICP has significantly higher number of fixations than DED, RMFD and LMFD. Pilots have to be able to 'see and process' the information to understand the situation, and then, to 'project' the situation in the near future. There is a long-standing argument concerning bottom-up or top-down visual processes in the eye movement literature. It is observed in this research that pilots applied both bottom-up and top-down visual processes, depending on the salience of information or previous experience.

Keywords: Aviation Safety, Eye Movement, Cognitive Processes, Fixation.

* Corresponding author.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 325–334, 2014.

1 Introduction

Eye scan pattern is one of the methods for assessing a pilot's cognitive process in the cockpit based on physiological measures (Ayazet al, 2010). It can provide numerous clues concerning the mental process of encoding information perceived by pilots by using in-flight visual behaviors, such as what areas of interest (AOIs) they scan, dwell and attend (Salvucci and Anderson, 1998). Eye movement can be measured continuously and objectively as these are able be recorded without interrupting pilot's activities. The visual information captured by eye tracking tools provides for the possibility of eye movement fluctuations while operating the task in hand occurring over short time intervals (Ahlstrom and Friedman-Berg, 2006). One more advantage is that eye movements are a sensitive and automatic response which may serve as a window into the process of the SA mechanism and reflection the mental state of a pilot (Kuo, Hsu and Day, 2009). For example, gaze trajectories can indicate pilot's attention distribution when he or she encounters certain displays of the cockpit interface or outside, such as terrain and direct fixation on specific AOIs in real time (Henderson, 2003; Pomplun and Sunkara, 2003). However, eye tracking technologies still have their limitations. For instance, the point the pilot fixated upon is not definitely where the attention was accurately located, which is known as "look but didn't see" (Shinar, 2008). SA has been recognized an essential component within a pilot's cognitive process in the domain of aviation (Sohn and Doane, 2004). Endsley (1995) defines three levels of SA which is linked closely with the major components within cognitive processes. The first level is to perceive environmental cues, such as warning lights in the cockpit. The second level is a process of comprehending the cues based on knowledge and experience. The third level is to predict the possible situation in the near future and project the related measurements to resolve the specific status.

There were considerable arguments regarding gaze control theories for decades: bottom-up and top-down visual processes (Henderson, 2003). There is an increasing need for further investigation of the relationship between gaze control and SA performance. The bottom-up visual process is stimulus-based and generated from the saliency of environment. It can be explained by level one of SA: perception of the cues; on the other side, the top-down approach is a knowledge-based theory that directs by internal cognitive process. The gazes are controlled to see a specific AOI to acquire the information to satisfy the task in hand. It complied with the three levels of situational awareness theory proposed by Endsely (1995). Pilot has to perceive the stimulus in the cockpit, understand the encountering situation, and predict the possible consequences. These visual searching within a flight deck are critical for collecting information, and over 75% of pilot errors are caused by perceptual failures (Jones and Endsley, 1996). It highlights the importance of the study concerning gaze control and SA performance. However, the empirical study of gaze control in aviation is relatively scarce compared with other eye movement behaviors such as fixation or saccade, not to mention the research of gaze control and three-level of SA. The visual behaviour directing gaze points to a specific AOI is attracted by the salient stimulus or controlled by a pilot's intention. Previous researches (Bellenkes, Wickens and Kramer, 1997; Ratwani, McCurry and Trafton, 2010) emphasize on how much fixation or

how long the duration on an AOI is held approximately stable on the fovea of the retina to identify whether pilot's visual behaviour is meaningful.

Pilots need to allocate attention to the interior and exterior of the cockpit to collect information and make decisions (Janis et al., 1996). However, pilots make more errors under stress such that attention tends to be focused on central information to the neglect of peripheral cues, resulting in tunnel vision (Orasanu, 2005). It was found that peripheral vision is useful for detecting objects, especially essential for detecting moving objects outside the fovea (Yang, 2012). Moreover, pilots with different levels of flying experience show various patterns of the usage of peripheral vision. More experienced pilots are more likely than the less experienced to use peripheral vision to process a wider field of visual cues, allowing experienced pilots to perform the main task while still obtaining the needed information (Kasarskis et al., 2001). Therefore, from an information processing perspective, the capability of peripheral vision is associated closely with cue acquisition and cue interpretation, which can be an index to evaluate pilot's SA performance that enables task-related information to be engaged and the problem to be resolved (Wiggins, 2006). Recognizably, peripheral vision is also linked closely with the bottom-up visual process, but it is impacted by the initial fixation location becoming longer and less gaze moving around the operational environment (Jungkunz and Darken, 2011). Also, pupil size is significantly influenced by the factor of task difficulty, and it is relevant to the operator's cognition loading. However, it is very complicated to interpret due to the influence from multiple factors such as cognitive workload, context complexity, environmental illumination and gaze angle (Pomplun and Sunkara, 2003; Gabay, Pertzov and Henik, 2011).

By utilizing a combination of an eye tracking device and flight simulator, pupil size can be collected for further analysis of pilots' cognitive processes in terms of attention allocation and SA performance at certain phase of flight operations, and this can be correlated with training and evaluation in aviation. This study combines an F-16 flight simulator and portable eye tracking device to investigate pilots' visual scan pattern and SA performance during an air-to-surface mission If the relationship of gaze points, fixation, pupil size and perceived workload related to SA performance could be identified in flight operations, then eye tracking tools could be considered for use in combination with flight simulators to improve training efficiency in the future.

2 Method

2.1 Subjects

There are 20 participants of F-16 pilots. The ages of participants are between 26 and 46 years old, and the total flying hours are between 400 and 3,250 hours.

2.2 Apparatus

Flight Simulator. The F-16 flight simulator is high-fidelity training device. It utilizes an actual cockpit with identical display panels, layout and controls to those in the actual

aircraft. This simulator provides a realistic representation of the flight management systems. The instructors can observe the pilot's performance via three screens without any intrusion. The scenario is designed to replicate an air-to-surface task. It is a challenging situation for subjects to perform as it represents a high demand flying task combined with hostile threats. Subjects not only have to execute the task precisely by operating the aircraft, but also have to follow navigation system entering the appropriate codes by using various flight deck interfaces. Simultaneously, subjects have to intercept the proper route and turn toward the target at an altitude of 500 feet with speed of 500-KIAS (Knots Indicated Air Speed), then performing a steep pop-up maneuver to increase altitude abruptly for appropriate reconnaissance, following by dive and roll-in toward the target to avoid hostile radar lock-on. When approaching the target, subjects have to roll-out, level the aircraft, aim at the target, release the weapon, and finally pull-up with a 5～5.5 G-force to break-away from the range.

Eye Tracking Device. Pilot's eye movements were recorded using a mobile head-mounted eye tracker (ASL Series 4000) which is designed and built by Applied Science Laboratory. It is light (76 g) and portable meaning it is easy for subjects to move their head without any limitations during the air-to-surface maneuvers. Video records the pattern of eye movements and the related data were collected and stored using a Digital Video Cassette Recorder (DVCR) and then transferred to a computer for further processing and analysis. The sampling frequency for eye movements was 30 Hz. The definition of an eye fixation point was when three gaze points occurred within an area of 10 by 10 pixels with a dwell time which was the time spent per glance at a location. There were five AOIs set up to collect subjects' eye movement data. Those AOIs were selected for performing the task of air-to-surface. AOI-1: Head-up Display (HUD); AOI-2: Integrated Control Panel (ICP); AOI-3: Data Entry Display (DED); AOI-4: Right Multiple Function Display (RMFD); and AOI-5: Left Multiple Function Display (LMFD).

2.3 Research Design

All subjects undertook the following procedures, (1) completed the demographical data including training experience and total flight hours (5 minutes); (2) a briefing of the study and the air-to-surface scenario (10 minutes); (3) calibration of the eye tracking device by using three points distributed over the cockpit display panels and screen (10-15 minutes); (4) participants performed the air-to-surface task (3-5 minutes).

3 Results

Twenty qualified mission-ready F-16 pilots participated in this research. The ages of participants are between 26 and 46 years old (M=33, SD=6); total flying hours between 400 and 3,250 hours (M=1358, SD=882); F-16 type flying hours between 101 and 2,270 hours (M=934, SD=689). Subjects' eye movement data described by number of fixation and number of gaze points are shown as table 1; subjects' average fixation duration and pupil diameters in five AOIs are shown as table 2.

There were significant differences in the pilots' number of gaze points among five different AOIs, F (4, 95) = 533.84, p<.001, η2ρ = .97. Further comparisons using post-hoc Bonferroni adjusted tests showed HUD has a significantly higher numbers of gaze points than ICP, DED, RMFD and LMFD; and ICP has significantly higher gaze points than DED, RMFD and LMFD. Also, there were significant differences in pilots' number of fixation among five different AOIs, F (4, 95) = 306.98, p<.001, η2ρ =.94. Further comparisons using post-hoc Bonferroni adjusted tests showed HUD has significantly higher number of fixation than ICP, DED, RMFD and LMFD; and ICP has significantly higher number of fixations than DED, RMFD and LMFD (table 1).

Table 1. Subjects' Eye Movement data for Number of Fixation and Gaze Points

Subject	Age	Total hours	Number of Gaze Points					Number of Fixations				
			HUD	ICP	DED	RMFD	LMFD	HUD	ICP	DED	RMFD	LMFD
1	27	550	2778	52	0	5	3	386	4	0	0	0
2	30	630	2307	420	20	28	5	297	27	2	3	0
3	41	2186	2187	67	4	4	5	321	6	0	0	0
4	26	400	2350	109	9	0	0	256	11	0	0	0
5	30	550	3044	191	71	2	0	457	24	11	0	0
6	28	620	2775	156	10	1	17	393	13	0	0	2
7	28	630	3432	76	69	3	1	456	10	12	0	0
8	35	1300	2833	117	1	51	0	447	7	0	4	0
9	35	1500	2339	99	1	0	4	312	12	0	0	0
10	42	3250	2925	60	16	37	1	421	5	1	6	0
11	28	582	2879	47	16	10	1	427	3	2	2	0
12	31	1000	1726	158	32	2	0	196	21	2	0	0
13	31	1032	2390	105	1	6	0	339	11	0	0	0
14	37	1650	1948	262	0	24	12	227	28	0	3	2
15	41	1900	3560	104	7	0	20	531	13	0	0	3
16	27	600	2563	55	4	16	0	297	3	0	2	0
17	37	1500	2925	16	13	57	0	427	1	2	8	0
18	34	1458	2086	24	1	0	0	296	1	0	0	0
19	46	2800	2155	119	102	6	7	269	8	9	0	0
20	41	3030	2365	70	0	3	0	330	3	0	0	0
M	33.75	1358.4	2578.35	115.35	18.85	12.75	3.80	354.25	10.55	2.05	1.40	0.35
SD	6.04	882.94	478.28	92.80	28.53	17.51	5.94	88.89	8.42	3.83	2.33	0.88

HUD: Head-up Display; ICP: Integrated Control Panel; DED: Data Entering Display; RMFD: Right Multiple Function Display; LMFD: Left Multiple Function Display

There were significant differences in pilots' average fixation among five different AOIs, F (4, 95) = 21.04, p<.001, η2ρ = .53. Further comparisons using post-hoc Bonferroni adjusted tests showed HUD has a significantly higher numbers of gaze points than DED, RMFD and LMFD; and ICP has significantly higher gaze points than DED, RMFD and LMFD. Also, there were significant differences in pilots' pupil diameter among five different AOIs, F (4, 95) =10.42, p<.001, η2ρ = .35. Further comparisons using post-hoc Bonferroni adjusted tests showed HUD has a significantly larger pupil diameter than DED and LMFD; and ICP has significantly larger pupil diameter than DED and LMFD (table 2).

Table 2. Performance, Workload and Average of Fixation Duration and Pupil Diameter

Subjects	Average Fixation Duration (ms)					Average Pupil Diameter in Region(pixel)				
	HUD	ICP	DED	RMFD	LMFD	HUD	ICP	DED	RMFD	LMFD
1	140	150	0	0	0	88.34	86.75	0	84.76	87.41
2	140	120	130	130	0	79.02	71.76	70.28	73	0
3	140	130	0	0	0	89.35	91.61	87.06	80.79	87.99
4	130	130	0	0	0	91.86	89.81	65.49	0	0
5	160	140	150	0	0	81.47	82.65	74.83	85.53	0
6	150	140	0	0	180	95.5	99.05	66.72	0	98.02
7	140	150	150	0	0	81.58	84.3	84.65	0	0
8	140	140	0	180	0	65.26	64.91	67.54	70.77	0
9	130	140	0	0	0	71.71	73.21	0	0	74.75
10	140	150	130	140	0	111.47	111.62	104.82	106.12	0
11	150	130	170	100	0	82.46	83.4	76.04	75.1	0
12	140	140	150	0	0	104.98	104.13	100.37	101.81	0
13	150	130	0	0	0	86.8	87.96	0	82.14	0
14	130	140	0	140	130	105.34	102.7	0	104.02	110.09
15	150	150	0	0	180	74.95	76.09	78.12	0	74.98
16	130	110	0	130	0	100.52	102.63	104.03	96.02	0
17	140	170	130	140	0	76.72	78.45	73.8	77.95	0
18	140	170	0	0	0	73.59	74.04	0	0	0
19	130	130	120	0	0	71.49	76.77	69.31	72.85	74.95
20	140	130	0	0	0	88.28	87.48	0	87.02	0
M	140.50	139.50	56.50	48.00	24.50	86.03	86.47	56.15	59.89	30.41
SD	8.26	14.68	71.69	68.41	60.57	12.67	12.51	39.43	41.44	43.19

4 Discussion

There are an average 2,578 number of gaze points on the HUD, however, there are only 354 fixations recorded by the eye tracking device (table 1). The setting of fixation in this study is that three gaze points occurred within an area of 10 by 10 pixels with the time spent per glance at allocation. The gaze control is the process of directing fixation through area of interests in the service of on-going perceptual, cognitive and behavior activities which are important for pilots to seek task relevant information. Fixation point is meaningful and is closely linked to attention allocation, however, gaze point is the foundation of fixation and it triggers pilots shifting attention to different AOIs in order to perform multiple tasks simultaneously, such as searching for target, keying data, analyzing information, and operating the aircraft to complete the mission. There is a close relationship between peripheral vision and gaze points to be observed. While pilots rapidly shift gazes from buttons within the ICP interface, their fingers can precisely key-in a series of codes without forming a fixation, and simultaneously search for the outside target. It demonstrates that gaze might be the precursor of fixation and enable the peripheral vision processing information promptly.

Previous research on gaze control has focused on two potential approaches; bottom-up of stimulus-based information generated from the image, and top-down of memory based knowledge generated from internal visual and cognitive systems (Henderson, 2003).There was an argument concerning bottom-up or top-down visual processes on the eye movement researches for a long time, it is observed by this

research that pilots applied both bottom-up and top-down visual processes depending on the prominence of information or previous experience. The top-down visual process indicates that the pilot recognized the subsequent engagement and planned the tactical strategies of air-to-surface by inputting navigation data into the ICP interface. Pilots have to move their fixations shifting to the buttons of ICP in order to guide his fingers to the specific number. When the directing attention allocation is completed, pilots relocate their fixations to the DED to determine if the information is precisely displayed. The bottom-up eye movement explains that the salient cues attract pilots' gazes to the objects by conducting a visual scan to perceive the unusual signal, such as the pilot moving gazes from surface target to the activated warning light on the HUD, reset on the master caution, then continued to aim at the surface target to complete the task. The analysis of frame-by-frame DVCR data of the eye tracking device found pilots also applied top-down visual process in the air-to-surface task. The integration of bottom-up and top-down visual processes might explain the three-levels of SA model as described by Endsely (1995); pilots perceived the warning light (level-1) and realized which system was malfunctioning (level-2), then predicted the malfunction's impact to the task (level-3). In this study, the level-1 of SA is a bottom-up approach for perceiving the stimulus of an activated warning light, level-2 and level-3 are top-down visual processes for understanding the stimulus by cross-checking the information from the HUD and relevant AOIs, then projecting the future situation by entering the codes to ICP for conducting the tactical manoeuver.

There are 94% of pilots' gaze points and 96% of fixations on the HUD, whilst performing the air-to-surface task. Although pilots have to key different codes into the ICP for aiming and releasing the weapon to target, it represents only 3% of fixation on the ICP. This phenomenon can be observed by analyzing eye tracking DVCR data which shows that while pilots are keying the codes into the ICP, they are also simultaneously searching for the surface target. To complete the task, pilots have to prioritize and switch attention between different AOIs depending on the specific stage of operating requirements for keying the navigation data. The LMFD mainly provides a moving map with terrain, while the pilots' priority information is altitude, speed and vertical speed whilst the target on the surface is in sight. It explains the low number of gaze points and fixation on the LMFD recorded in the air-to-surface task. Furthermore, blinking might reduce the number of gaze points and number of fixations counted, as it is an involuntary act of shutting and opening the eyelids which blocks the pupil and cornea from the illuminator resulting in raw data points missing. Searching for information in the cockpit and aiming at targets involve pilots' attention allocation. Pilots have to be able to 'see and process' the information to understand the situation, and then, to 'project' it in the near future (Endsely, 1995). It is a series of cognitive processes that constitute pilot aeronautical decision-making (ADM).

A close relationship between peripheral vision and gaze points can be observed as pilots rapidly shift gazes from buttons within ICP, while their fingers precisely key-in a series of codes without forming a fixation. Pilots not only have gaze points on the buttons of the ICP for entering a series of codes, but also simultaneously search for the outside target. In this study, pilots have the average of 2,578 gaze points, however, the average of pilots' fixation number were only 354 recorded by the eye tracker. This

finding supports previous research by Henderson (2003) that the gaze control is an important topic in scene perception for seeking out task-relevant visual information and allocating attention. It provides evidence that gaze might be the precursor of fixation and enable peripheral vision in processing information promptly. According to the definition of fixation in this research, three gaze points occurred within an area of 10 by 10 pixels with a glance. Fixation point is definitely meaningful and is closely linked to attention allocation (Ratwani, McCurry and Trafton, 2010). However, gaze point is the foundation of fixation and it triggers pilots shifting attention to different AOIs whilst performing multi-tasks simultaneously, such as searching information, keying information, analyzing information, and operating the aircraft.

Research has shown that the retina needs about 80 ms of seeing a new image before that image is registered in normal light conditions. This doesn't mean that pilots consciously have noticed any change; it is only that the eye has registered a change. Furthermore, it has been observed that seeing a word in order to perceive it needs between 50-60 ms, while looking at a picture might need more than 150 ms to be able to interpret the content. The average fixation duration on the HUD and ICP are significant higher than DED, RMFD and LMFD. The information can be identified rapidly within the duration of single fixation, but this rapid apprehension may require attention allocation. The average of fixation duration on the HUD and ICP are 140 ms in this research (table 2), which differs from previous research where the overall average fixation duration was approximately 400 ms on the Primary Flight Display (PFD) and Navigational Display (ND) (Diez et al., 2008). The difference might be that the contexts of the research are different; one in a civil aviation setting, the other in military tactical operations. Generally, military pilots have higher standard of response time (shorter) compared with civil pilots, as the tactical operation has to be precisely accomplished under time pressure. Therefore, military pilots have shorter average fixation duration than civil pilots.

Pupil size is affected by human emotional and cognitive processes, and the increase in pupil size is an indicator of cognitive load (Bee et al., 2006). Under conditions of controlled illumination in the training simulator, pupil size is an effective and reliable measure of mental workload, as pupil size can reveal the condition of cognitive load, and the increases in pupil size correlate with increases in mental workload. Table 2 shows that pilots' pupil size at the ICP is the largest, followed by HUD and RMPD, LMFD is the smallest for the pupil size. When approaching the target, pilots have to roll-out, level off the aircraft, and with only very limited time to aim at the target, release the weapon and pull-up with a 5~5.5 G to break-away from the range, otherwise the aircraft will be exposed to high risk. Pilots conduct lots of tactical manoeuvers to level-off the aircraft under hostile conditions and with limited time to aim at the target. If they cannot successfully aim and lock on the target, the mission has failed. All the critical information related to mission completion were provide by HUD and ICP; HUD shows all the important navigation and weapons information such as pitch, bank, airspeed/Mach, heading, altitude, horizon line, load factor, navigation information, air-surface target information, and the ICP is used for weapons release, landing, NAV/COM frequencies and to show air or surface target information. It is the reason

why the pupil size on the ICP and HUD were significant larger than DED and LMFD (table 2).

The number of fixations multiplied by average fixation duration is the total fixation duration. Pilots have large amounts of total fixation duration time on the HUD. This demonstrates a phenomenon of focusing on particular parameters on the HUD which might potentially result in tunnel vision or overlooking critical parameters and missing the target. A limitation of simulator training is that the instructor cannot identify which AOIs a trainee is looking at to get information during training. If a trainee's real-time visual scan pattern can be recorded and displayed on the control panel simultaneously for an instructor to be aware of their attention allocation, it might improve training effectiveness and therefore also pilots' performance and aviation safety.

5 Conclusion

It is very important to improve military pilots training for the air-to-surface task, as it is the training element with highest risk of control flight into terrain (CFIT). Understanding a pilot's visual scan pattern and attention distribution during the air-to-surface task will allow aviation professionals to develop effective training. This research observed that over 90% of pilots' gaze points and fixations are on the HUD. It implies that the HUD might provide all the necessary information for pilots to perform the air-to-surface task; or it might be the evidence of pilots' over-reliance on the HUD. Therefore, the intervention of training could focus on the HUD to address how to improve the function of HUD, or how to conduct proper attention allocation between AOIs. The limitation of traditional simulator training is that there is no specific feedback of a trainee's visual scan pattern provided to the instructor to address the critical timing of attention distribution on the flight deck. This is because a pilot's visual scan patterns and attention allocation could not be observed simultaneously by an instructor. Eye tracking devices can aid in capturing a pilot's attention allocation where traditional flight simulators training were lacking. Therefore, a simulator integrated with eye tracking devices will be a creative method to promote safety and effectiveness in flight operations.

References

1. Ahlstrom, U., Friedman-Berg, F.J.: Using eye movement activity as a correlate of cognitive workload. International Journal of Industrial Ergonomics 36(7), 623–636 (2006)
2. Ayaz, H., Willems, B., Bunce, B., Shewokis, P.A., Izzetoglu, K., Hah, S., Onaral, B.: Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. In: Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations, pp. 21–31 (2010)
3. Bee, N., Prendinger, H., Nakasone, A., André, E., Ishizuka, M.: AutoSelect: What you want is what you get: Real-time processing of visual attention and affect. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Weber, M. (eds.) PIT 2006. LNCS (LNAI), vol. 4021, pp. 40–52. Springer, Heidelberg (2006)

4. Bellenkes, A.H., Wickens, C.D., Kramer, A.F.: Visual scanning and pilot expertise: The role of attentional flexibility and mental model development. Aviation, Space, and Environmental Medicine 68(7), 569–579 (1997)
5. Cannon-Bowers, J.A., Salas, E., Pruitt, J.S.: Establishing the boundaries of a paradigm for decision-making research. Human Factors 38(2), 193–250 (1996)
6. Diez, M., Boehm-Davis, D.A., Holt, R.W., Pinney, M.E., Hansberger, J.T., Schoppek, W.: Tracking pilot interaction with flight management systems through eye movements. In: Proceeding of the Human Factors and Ergonomics Society 52nd Annual Meeting (2008)
7. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors: The Journal of the Human Factors and Ergonomics Society 37, 32–64 (1995)
8. Kuo, F.Y., Hsu, C.W., Day, R.F.: An exploratory study of cognitive effort involved in decision under Framing-an application if the eye-tracking technology. Decision Support Systems 48, 81–91 (2009)
9. Henderson, J.M.: Human gaze control during real-world scene perception. TRENDS in Cognitive Sciences 7(11), 498–504 (2003)
10. Jones, D.G., Endsley, M.R.: Sources of situation awareness error in aviation. Aviation Space and Environmental Medicine 67, 507–512 (1996)
11. Jungkunz, P., Darken, C.J.: A computational model for human eye-movements in military simulations. Computational and Mathematical Organization Theory 17(3), 229–250 (2011)
12. Kasarskis, P., Stehwien, J., Hickox, J., Aretz, A., Wickens, C.: Comparison of expert and novice scan behaviors during VFR flight. In: The 11th International Symposium on Aviation Psychology, Columbus, OH (2001)
13. Orasanu, J.: Crew collaboration in space: a naturalistic decision-making perspective. Aviation Space and Environmental Medicine 76(6) (suppl.), B154–B163 (2005)
14. Pomplun, M., Sunkara, S.: Pupil dilation as an indicator of cognitive workload in human-computer interaction. Paper Presented at the Proceedings of the International Conference on HCI (2003)
15. Salvucci, D.D., Anderson, J.R.: Tracing eye movement protocols with cognitive process models. In: Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, pp. 923–928. Lawrence Erlbaum Associates, Hillsdale (1998)
16. Gabay, S., Pertzov, Y., Henik, A.: Orienting of attention, pupil size, and the norepinephrine system. Attention Perception Psychophysics 73, 123–129 (2011)
17. Ratwanti, R.M., McCurry, J.M., Trafton, J.G.: Single operator, multiple robots: An eye movement based theoretic model of operator situation awareness. In: Proceedings of the Fifth ACM/ IEEE International Conference on Human-Robot Interaction, pp. 235–242. Nara, Japan (2010)
18. Shinar, D.: Looks are (almost) everything: where drivers look to get information. Human Factors: The Journal of the Human Factors and Ergonomics Society 50(3), 380–384 (2008)
19. Sohn, Y.W., Doane, S.M.: Memory processes of flight situation awareness: Interactive roles of working memory capacity, long-term working memory, and expertise. Human Factors: The Journal of the Human Factors and Ergonomics Society 46(3), 461–475 (2004)
20. Wiggins, M.W.: Cue-based processing and human performance. Encyclopedia of Ergonomics and Human Factors, 641–645 (2006)
21. Yang, J.H., Huston, J., Day, M., Balogh, I.: Modeling Peripheral Vision for Moving Target Search and Detection. Aviation, Space, and Environmental Medicine 83(6), 585–593 (2012)
22. Zakowski, S., Hall, M.H., Baum, A.: Stress, stress management, and the immune system. Applied and Preventive Psychology 1, 1 (1992)

The Evaluation Model of Psychological Quality for Civil Aviation Student Pilot Based on Fuzzy Comprehensive Evaluation

Shu Li[1] and Yang You[2]

[1] Flight Technology College, Civil Aviation University of China (CAUC), Tianjin, China
lishu-smile@hotmail.com
[2] Flight Training Management Division,
Civil Aviation University of China (CAUC), Tianjin, China
yyou@cauc.edu.cn

Abstract. To establish the civil aviation student pilot psychological quality evaluation model, investigation was carried out about the psychological quality indicator system. This paper analyzed and summarized the relevant research literature, extracting the psychological quality indicators which effect flight training performance. 20 experts identified 4 categories of 21 psychological quality evaluation indicators, and established the civil aviation student pilot psychological quality evaluation indicators system. Using Delphi's analysis, the weight of each indicator was determined. The flight psychological quality evaluation model was constructed using fuzzy comprehensive evaluation method and the corresponding internet -based questionnaires were sent out to 100 student pilots. The evaluation results were compared with their flight training results, verifying the correctness and validity of the model indicators. The evaluation model can play an important role in psychological selection and psychological training for student pilots, which could further reduce the grounded rate and avoid unnecessary losses.

Keywords: civil aviation student pilot psychological quality, psychological quality indicator system, fuzzy comprehensive evaluation, psychological quality evaluation model.

1 Introduction

Flight safety has always been the focus of the civil aviation industry worldwide, and the key factors that affect flight safety is the human factors. Statistics show that flight accidents caused by human factors account for about 75%, and the psychological factors constitute the major part of the human factors. The aviation developed countries and regions in Europe and America have researched and established a set of psychological management system suited to their national pilots' selection, training and evaluation and the number of commercial transport pilots being eliminated because of psychological reasons are far more than physical reasons. Currently in China, flight accidents caused by health reasons have been fewer and fewer due to the pilot's

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 335–343, 2014.

selection procedure, regular physical examination, and health assessment. But flying career particularity requires pilots have excellent psychological quality. China's civil aviation don't have uniform requirement about psychological selection and evaluation. Currently companies and civil aviation colleges' selection can only ensure physical examination, there is no uniform practice for psychological evaluation, and even some small companies do not conduct psychological evaluation. Now in China, civil aviation flight training academies have no psychological quality assessment and training systems, also lack of psychological quality continuous monitoring system.

Currently, limited by the amount of CCAR-141 flight training academies, the Chinese airlines annually enroll more than 3,000 student pilots and about two-thirds of them were sent to 32 foreign flight training academies. Subject to the region restrictions, it's difficult to carry out systematic psychological quality monitor. Take Civil Aviation University of China for example, 204 of the 252 total undergraduate student pilots were sent abroad in 2012. The student pilots have to face the risk of being grounded at any time during the theoretical study and flight training process, the grounded rate is around 15%-20%, which give them a lot of psychological pressure and make their flight efficiency decreased, their training progress slower, or even be grounded.

This paper will build a psychological quality evaluation model which could be feasible and convenient for evaluating and monitoring student pilots abroad. Timely guidance and help will be given to those whose is found in adverse psychological state, so that to reduce the grounded rate due to psychological reasons.

2 Civil Aviation Student Pilot Psychological Quality Evaluation Indicator System

Establishing appropriate indicator system is the key to evaluate student pilot psychological quality, related to whether we can fully reflect their real psychological condition, and thus directly affect whether the student pilot needs to be helped to adjust his psychological state for the flight training.

Based on 58 related research literature，227 related evaluation indicators were collected and analyzed, and 79 indicators were kept. An expert group composed of 5 psychology experts, 5 flight theory instructor, 5 flight instructors and 5 outstanding student pilots who have completed flight training identified 4 categories of 21 psychological quality evaluation indicators.

2.1 Psychological State

Foreign aviation training academies usually arrange a lot of flight training mission when the weather and control condition allows in order taking full advantage of the good weather and airspace. Therefore, in many cases, training mission will be extreme imbalance; training time is not fixed or regular, which greatly increased the psychological pressure on student pilots.

Psychological State is the indicator to reflect daily psychological condition of student pilot, which directly affect their behavior. There are six secondary indicators: No Pessimism, No Procrastination, No Depression, No Anxiety, No Sensitive, No Stress, and No Training Burnout.

2.2 Basic Ability

Before flight training student pilots need study theoretical knowledge first and pass various theoretical exams in country where they are trained. Most training academies have strict limits for the number of make-up, if cannot pass the exam within the specified time and limits, the student pilot will be suspended or even permanent grounded, resulting in some psychological pressure on them.

Basic Ability is the indicator to reflect if the student's ability is competent for study, which directly affect student's learning efficiency. There are four secondary indicators: Making Decision, Judgment and Reasoning, Communication and Coordination, Quantitative Relationship.

2.3 Smooth Flight

Every flight is a test. To ensure a smooth flight a good personality is an important factor affecting the state of the pilot work.

Smooth Flight is the indicator to reflect whether student pilot could stably display personal ability. There are six secondary indicators: Rigorous, Systematic, Self-discipline, Responsibility, Teamwork, and Emotional Stability.

2.4 Crisis Response

In flight training, student pilots are likely to encounter some unexpected events, which test if they can maintain calm and objective attitude to withstand the psychological pressure, and decisively solve the problem.

Crisis Response is the indicator to reflect if student pilots can handle crisis situations. There are five secondary indicators: Strong-willed, Affordability, Objective and Rational, Aggressive, Decisive.

3 Evaluation Model Based on Fuzzy Comprehensive Evaluation

Fuzzy comprehensive evaluation method is based on the theory of fuzzy mathematics and on the basis of fuzzy relations with their synthetic operations, with the help of function in fuzzy theory to express the state of the factors. It is a combination of qualitative and quantitative evaluation methods to solve a variety of factors on the material being evaluated. It is particularly suitable for comprehensive evaluation of the project that lots of indicators are difficult to quantify. The steps of fuzzy comprehensive evaluation method are as follows:

- Determine the factors U of psychological quality evaluation of distribution network. Each of these factors is represented by a single review. Supposing there are m reviews, and they constitute a reviews of discourse U .

$$U = \{u_1, u_2, ..., u_i, ..., u_m\} \tag{1}$$

i = (1, 2, …, m), u_i is one of the review.

- Establish the quality level V of the reviews. In accordance with the degree of pros and cons, divide u_i into n levels, that is the level of discourse V .

$$V = \{v_1, v_2, ...v_j, ... v_n\} \tag{2}$$

j = (1, 2, …, n), v_j is one of the levels.

- Establish affiliation and obtain fuzzy evaluation matrix R. Every u_i has a membership for every v_j . Usually the score of the evaluation factors are given by experts. Then the fuzzy evaluation matrix R is:

$$\mathrm{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix} \quad 0 \le r_{ij} \le 1 \tag{3}$$

$R_i = (r_{i1}, r_{i2}, \ldots r_{ij}, \ldots, r_{in})$ is one of the fuzzy evaluation vectors.

- Indicators of different weights may lead to different evaluation results. The traditional methods to determine the weight is expert assignment method, using the expert investigation method to determine the subjective weight A .

$$A = (a_1, a_2, \cdots, a_n) \qquad \sum_{i=1}^{n} a_i = 1 \tag{4}$$

a_i is the weight which given by experts.

- According to the weights of all evaluation factors make the fuzzy comprehensive evaluation.

$$\mathrm{B} = \mathrm{A} \circ \mathrm{R} = (a_1, a_2, \cdots, a_m) \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix} = (b_1, b_2, \cdots, b_n) \quad (5)$$

Fuzzy vector B = (b_1, b_2, …, b_n) is the needed result.

4 The Evaluation Model of Psychological Quality for Civil Aviation Student Pilot

- Corresponding factor set and sub factors sets are as follows and shown in Table 1.

Table 1. Indicator System of Civil Aviation Student Pilot Psychological Quality Evaluation of Distribution Network

	Factor Set	Sub Factors Sets
Psychological Quality U	Psychological State U_1	No Procrastination u_{11}
		No Depression u_{12}
		No Anxiety u_{13}
		No Sensitive u_{14}
		No Stress u_{15}
		No Training Burnout u_{16}
	Basic Ability U_2	Making Decision u_{21}
		Judgment and Reasoning u_{22}
		Judgment and Reasoning u_{23}
		Quantitative Relationship u_{24}
	Smooth Flight U_3	Rigorous u_{31}
		Systematic u_{32}
		Self-discipline u_{33}
		Responsibility u_{34}
		Teamwork u_{35}
		Emotional Stability u_{36}
	Crisis Response U_4	Strong-willed u_{41}
		Affordability u_{42}
		Objective and Rational u_{43}
		Aggressive u_{44}
		Decisive u_{45}

- According to the psychological quality which possibly is contributed by each factor, we divide the psychological quality level into five levels, namely:

$$V = \{V_1, V_2, V_3, V_4 V_5\} = \{excellent, good, average, fair, poor\} \quad (6)$$

- The weights given by experts are as follows:

  A = (0.3, 0.2, 0.3, 0.2); A_1 = (0.2, 0.2, 0.1, 0.1, 0.2, 0.2); A_2 = (0.25, 0.25, 0.3, 0.2); A_3= (0.1, 0.2, 0.2, 0.2, 0.15, 0.15); A_4= (0.2, 0.2, 0.2, 0.2, 0.2).

- The fuzzy evaluation matrix's establishment uses the expert investigation method. The content which is going to appraise will be designed into the judge advice questionnaire table by the analysis staffs and use letters to distribute to 20 experts. According to their experience experts make clear judgments about the contents by an anonymous way. Depend on the percentage that some appraisal occupies all experts' appraisal number as the determination psychological factor degree of the psychological factors to establish psychological quality evaluation matrix.

Table 2. Civil Aviation Student Pilots Psychological Quality Fuzzy Evaluation Matrix

	V_1	V_2	V_3	V_4	V_5
u_{11}	0.2	0.15	0.5	0.15	0
u_{12}	0.2	0.5	0.2	0.1	0
u_{13}	0.1	0.4	0.2	0.2	0.1
u_{14}	0.2	0.3	0.4	0.05	0.05
u_{15}	0.2	0.2	0.4	0.1	0.1
u_{16}	0.3	0.45	0.2	0.05	0
u_{21}	0.3	0.5	0.15	0.05	0
u_{22}	0.2	0.2	0.4	0.1	0.1
u_{23}	0.4	0.4	0.1	0.1	0
u_{24}	0.1	0.3	0.3	0.2	0.1

	V_1	V_2	V_3	V_4	V_5
u_{31}	0.1	0.25	0.5	0.1	0.05
u_{32}	0.2	0.3	0.4	0.1	0
u_{33}	0.2	0.3	0.35	0.15	0
u_{34}	0.1	0.3	0.4	0.1	0.1
u_{35}	0.1	0.3	0.3	0.2	0.1
u_{36}	0.1	0.4	0.4	0.1	0
u_{41}	0.3	0.4	0.2	0.1	0
u_{42}	0.1	0.4	0.3	0.1	0.1
u_{43}	0.2	0.3	0.25	0.15	0.1
u_{44}	0.1	0.3	0.3	0.2	0.1
u_{45}	0.2	0.35	0.35	0.1	0

- Establish the fuzzy relation matrix for the single factor fuzzy evaluation as follows:

$$B_1 = A_1 \circ R_1 = (0.2,0.2,0.1,0.1,0.2,0.2) \circ \begin{pmatrix} 0.2 & 0.15 & 0.5 & 0.15 & 0 \\ 0.2 & 0.5 & 0.2 & 0.1 & 0 \\ 0.1 & 0.4 & 0.2 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.4 & 0.05 & 0.05 \\ 0.2 & 0.2 & 0.4 & 0.1 & 0.1 \\ 0.3 & 0.45 & 0.2 & 0.05 & 0 \end{pmatrix}$$

$$B_1 = (0.21, 0.33, 0.32, 0.105, 0.035)$$

$$B_2 = A_2 \circ R_2 = (0.265, 0.355, 0.2275, 0.1075, 0.045)$$

$$B_3 = A_3 \circ R_3 = (0.14, 0.31, 0.385, 0.125, 0.04)$$

$$B_4 = A_4 \circ R_4 = (0.18, 0.35, 0.28, 0.13, 0.06)$$

- According to the weights of all evaluation factors make the fuzzy comprehensive evaluation.

$$B = A \circ R = (0.3, 0.2, 0.3, 0.2) \begin{pmatrix} 0.21 & 0.33 & 0.32 & 0.105 & 0.035 \\ 0.265 & 0.355 & 0.2275 & 0.1075 & 0.045 \\ 0.14 & 0.31 & 0.385 & 0.125 & 0.04 \\ 0.18 & 0.35 & 0.28 & 0.13 & 0.06 \end{pmatrix}$$

$$B = (0.1940, 0.333, 0.313, 0.1165, 0.0435)$$

- According to the maximum degree of membership principle, take the evaluating indicator which the maximum membership degree B in the calculation correspond. Namely, it is the final evaluating consequence. With the formula expressed as:

$$V_b = \left\{ V_l \middle| V_l \to \max_{i=1}^{n} (b_i) \right\} \tag{7}$$

$$V = \{V_2\} = \{good\}$$

So the student pilot's psychological quality is good.

5 Conclusions and Prospects

5.1 Conclusions

According to the indicator system of civil aviation student pilot psychological quality evaluation, the evaluation factor sets, evaluation criteria sets, membership function and objective weight sets were established, and then a comprehensive fuzzy evaluation for the psychological quality of one student pilot was conducted, by using the principle of maximum method. The result matches the training performance of him.

An internet-based questionnaire of psychological quality test for student pilots had been developed according to the psychological quality evaluation model. The questionnaire, including total 274 questions, taking about 90 minutes, could be taken worldwide by using username and password on the website. The test report would be submitted to the experts and the analyst result would be sent to the individual taken the test.

After 100 student pilots being tested, the comparison of their results in the test and their routinely training performance (fast, normal, slow, facing grounded, grounded) showed that a positive relationship between the two results. The student pilots who got better evaluation in the test were more incline to make better performance in the training courses.

The average grounded rate of student pilots is 15% to 20%. The ground rate of student pilots who scored average and above in the psychological quality evaluation was 10.26% which is much lower than the 33.51% ground rate of the ones who scored worse than average in the evaluation. To sum up, the test results show the validity of the evaluation index system and model.

5.2 Prospects

The research results showed that the psychological quality evaluation model had reflected psychological quality of student pilots which infect their flying performance. Therefore this evaluation model could play an important role in psychological selection and psychological training for student pilots, by reducing ground rate and avoiding unnecessary losses.

In this research a civil aviation student pilot psychological quality evaluation model had been set up based on fuzzy comprehensive evaluation. However, the evaluation indicator system remains to be improved as no specific indicators of the research were in-depth analysis of ambiguity and the system validity is only qualitative analyzed without deeply quantitative detection of evaluation questionnaires and performance.

Seeking cooperation with international experts, to understand the needs of civil aviation student pilots' psychological quality, this research focused on finding new training methods of pilots' psychological quality, in order to improve the effect of pilots' psychological quality training. Establishment of the psychological quality archives of student pilots is crucial in continuing evaluation and instruction in the training process. Detailed records of pilots' psychological quality, not only greatly improve the psychological quality of the pilot, but could also provide a strong basis for flight training, and reduce ground ratio due to psychological causes.

Acknowledgement. It is a project supported by "Student Pilot Psychological Quality Continuous Assessment and Training System Research", which is the Class D Special Project of the Central University Basic Scientific Research (ZXH2012P005).

References

1. Civil Aviation Administration of China: AC-141-FS-2013-01R2. Advisory Circular (2013)
2. Walker, T.B., Lennemann, L., Doczy, E.: The Influence of Agility Training on Physiological and Cognitive Performance. Air Force Research Lab., Brooks (2010)
3. Justin, C., Michael, C., Steven, P.M.: Analysis of Personality Assessments as Predictors of Military Aviation Training Success. International Journal of Aviation Psychology 20, 92–109 (2010)
4. Du, Y.: Explore air traffic controller the psychological factors. Journal of Civil Aviation University of China 21, 69–71 (2003)
5. Zhiyong, X., Xudong, Z., Ming, Z., Fan, Y.: Application of ANP-based Multilevel Fuzzy Comprehensive Evaluation Methods to Post-evaluation for Grid Construction Projects. East China Electric Power 37, 488–491 (2009)
6. Aadms, R.J., Ericsson, A.E.: Introduction to cognitive processes of expert pilots. Hum. Perf. Extrem. Environ. 5, 44–62 (2004)
7. Shippmann, J.S., Ash, R.A., Battista, M.: The practice of competency modeling. Personnel Psychology 53, 703–740 (2000)
8. Hui, M., Shaowei, Y.: Application of Multi-level Fuzzy Evaluation Method in Route Plan Selection. Journal of Zhengzhou University 3, 134–138 (2009)

9. Hua, L., Yaohong, W.: The Application of Multi-level Fuzzy Comprehensive Evaluation Method in Social Assessment of Transport Investment Projects. Industrial Technology and Economy 10, 94–96 (2007)
10. Harper, R.P., Cooper, G.E.: Handling qualities and pilot evaluation. Journal of Guidance Control and Dvnamtis 9, 515–530 (1986)
11. Flin, R., Goeters, K.M., Martin, L., Hormann, H.J.: A generic structure of non-technical skills for training and assessment. In: 23rd Conference of European Association for Aviation Psychology, Vienna (1998)

A Study of the Relationship between Novice Pilots' Performance and Multi-Physiology Signals

Yanyu Lu, Jingjing Wu, and Shan Fu

School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai, China
luyanyu@sjtu.edu.cn

Abstract. The performance of pilots are related to the physiological response. In this study, we investigated the correlation between the pilot performance and physiological meters through recording the multiple physiological parameters in a flight. The results showed that in the flight phase from cutting off auto-pilot and auto-throttle, the performance of the subjects correlated strongly with heart rate and also correlated with fixation time or blink interval while the correlation between the performance and the saccade frequency or blink duration was weaker. When the subjects adjusted the flaps or controlled landing gear, the correlation between the performance and multiple physiological parameters, specially the cardiovascular parameters and fixation time, were more stronger compared with that at the other time.

Keywords: pilots' performance, physiological response, simulated flight, mental workload.

1 Introduction

Safety is one of the key element of civil aviation since the air transport system plays an important role in the world economic activity. The number of accidents decreased significantly as the level of technology and management increased, while human factors becomes more and more serious cause of the accidents. Therefore, to study and improve human factor is an important way to reduce accident rate and improve aviation safety.

An important aspect of human factor is assessment of workload since either excess workload or underload can degrade performance. Workload measurement has been used in a number of areas, such as military and industries. Workload measurement technologies can be classified into three categories: subjective ratings, physiological recordings and performance measures [1].

Subjective ratings include uni-dimensional and multi-dimensional scales. There are a lot of subjective measures applied in the different studies, such as modified Cooper-Harper scale, SWAT, and NASA-TLX [2]. Although subjective measures are easy to assign ratings and accepted by operators, it is difficult to compare workload on different task according to the rating scales.

Physiological recordings are based on the premise that the physiological responses in the body based on the nervous regulation and autoregulation are sensitive to the

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 344–350, 2014.

change of the workload [3]. The physiological measures are continual and objective measurement of the operate state, and not be affected by the subjective measures [4]. It is demonstrated that many physiological signals are sensitive to workload. The P300 EPR appears to respond to workload [5]. The cardiovascular response can reflect the aroused levels and be used in a variety of studies [6]. The eye blink measure is sensitive to the visual workload and eye movement can reflect the attention allocation and decision making [7].

Performance measures can be divided into primary-task and secondary-task measures. The primary-task measures consider the workload induced by the interested task, and can provide a direct indication of performance on the interested task. The secondary task is one that is subordinate to the primary task where dual tasks are required, and can provide a index of spare capacity. The rationale is that performance on the secondary task will decline as a function of the demands of the primary task [4].

Because of the different advantage of three kinds of workload measures, many researchers assessed or predicted pilots' workload though multiple measures. Jorna [8] reported the objective and subjective measurements on human interaction with air traffic control systems. Miyake [9] integrated physiological parameters and one subjective parameter through Principle Components Analysis into the multivariate workload evaluation index to assess the metal workload. Sonderegger and Sauer [10] examined the influences of situational factors on use behavior through taking performance data, subjective measures and physiological parameters. However, the relation between the pilots' performance and physiological response remains unclear. Pilots' performance depends on a number of factors: pilots' training levels, age, workload, their mood, etc. For the novice pilots with less training and skills, their performance is quite relative to the workload. In this study, we investigated the correlation between the pilot performance and physiological meters through recording the multiple physiological parameters in a flight.

2 Materials and Methods

2.1 Subjects

Six subjects with normal or corrected visual acuity (20/20; 3 males, 3 females, 20-30 years old) were recruited from Shanghai Jiao Tong University. All subjects were informed of the purpose and procedures of the experiments and signed an informed consent form prior to participation. The subjects had basic knowledge about aviation and were trained 8 times to familiarize with the flight in the simulator before the formal experiment. Before the experiment, they were given enough rest according to their mental and psychological status. At the end, they were given awards for their contribution to our research.

2.2 Apparatus

Simulator. The Boeing 777-200ER flight simulator was built up by using three workstations (CPU: 2.90GHz, 4GB DDR3 RAM, HP, China) and six 22-in LCD

screens (LE2201W, HP, China). The outside view of the flight simulator, which presented a real immersed experience of the flight to the subjects, was displayed on a semi-spherical screen with the diameter of 8 meters through three projectors.

Fig. 1. An Boeing-777 Flight Simulator

Physiological devices . An eye tracker (Smart Eye, Smart Eye AB, Sweden) was used to monitor and record pupil diameter, eye blink and movements signals. A physiological parameters monitoring equipment (Bio Harness, Zephry Technology, Annapolis, U.S.A) was wore by the subject to record the heart rate data, respiration amplitude and rate.

2.3 Procedure

The subjects were asked to perform the whole flight, including taking off, cruise and landing, which lasted about 10 minutes in the flight simulator with high fidelity. The subjects were required to take off from the runway of KJSC 30R. Before that, they had to set the auto-pilot and auto-throttle. After climbing to 2000 feet with maintaining level flight, the subjects opened the auto-pilot and auto-throttle according to the instruction, and then they only needed to supervise whether the flight was on the established route. They were required to disconnect the auto-pilot and auto-throttle and execute the landing before 10 nautical mile of the destination, and then they should maintain the aircraft about 3 to 5 degree of nose-up to descent and adjust the flaps at certain airspeed. When one dot from glideslope displayed, the subjects would hear an instruction from the experimenter to extract the gear, then they landed the aircraft on the runway. Therefore, this phrase was selected to investigate the relationship between the performance and the physiological response.

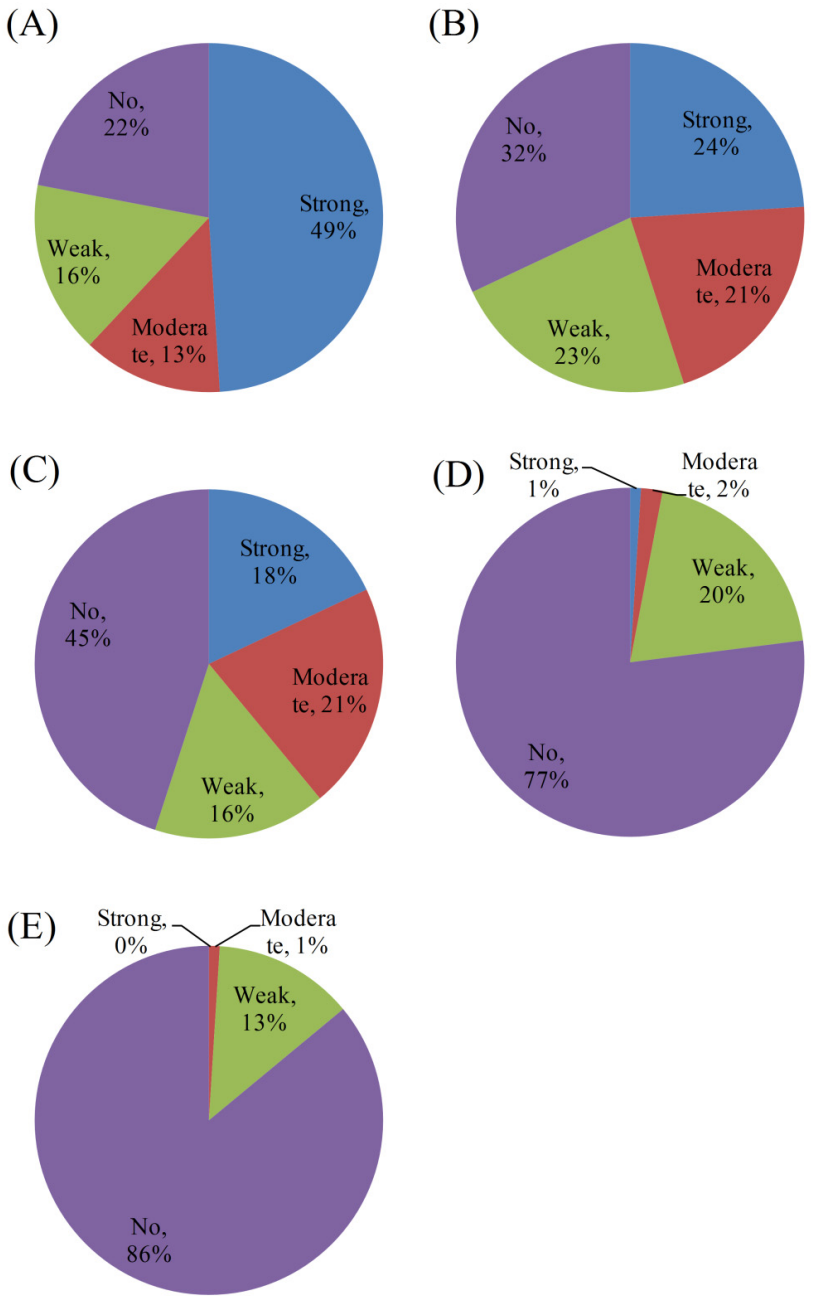


Fig. 2. The correlation between the performance and (A) heart rate, (B) fixation time, (C) blink interval, (D) blink duration and (E) saccade frequency. The correlation coefficient |R| >0.7: strong; 0.3<|R|<0.7: moderate; 0.3<|R|<0.1: weak; |R|<0.1: no.

2.4 Data Analysis

The performance, which represents the pilot's control, was integrated by the flight path deviation and acceleration. The physiological parameters contained heart rate, blink duration/interval, fixation time and saccade frequency. All the values recorded in the flight were normalized. The correlation between the performance and physiological parameters at each moment of the selected phase was analyzed using the Pearson chi-square test.

3 Results

The correlation analysis between the performance and physiological parameters in a simulated flight were shown in the results. In the flight phase from cutting off autopilot and auto-throttle, the performance of the subjects correlated with heart rate strongly (correlation coefficient > 0.7) for about 50% sampling sites and moderately (correlation coefficient: 0.3-0.7) for 13% sampling site; the correlation coefficient between the performance and fixation time or blink interval was more than 0.3 for about 30%-50% sampling site while the correlation between the performance and the saccade frequency or blink duration was weaker.

Taking into account the performance referring to the control, we analyzed the correlation at the moment when the subjects had a control action. When the subjects adjusted the flaps or controlled landing gear, the correlation between the performance and multiple physiological parameters, specially the heart rate, fixation time and blink interval, were more stronger compared with that at the other time.

Table 1. The correlation coefficient between the performance and multiple physiological parameters at different control action

	Heart rate	Fixation time	Blink interval	Blink duration	Saccade frequency
Flap1	0.984	0.872	-0.909	-0.268	-0.652
Flap2	-0.749	-0.841	0.476	0.536	0.589
Flap3	0.965	-0.515	-0.346	-0.006	0.103
Flap4	0.990	0.734	-0.897	0.513	-0.448
Landing	0.898	0.464	-0.506	0.760	-0.753
Others	0.094	-0.413	-0.519	0.331	0.417

4 Discussion and Conclusion

The results showed that multiple physiological parameters were correlated with the performance, especially the heart rate, blink interval and fixation time. As we know, heart rate is related with the variation of the emotional states of the human. It has been demonstrated that the change of heart rate can reflect the change of the workload, task difficulty, and human emotion. Moreover, it is well knew that the ocular motor

behavior was linked to the attention or cognition load [11]. It was proved that the increased subjective effort, and behavioural and physiological costs could improve the performance [12, 13], which are consistent with our results.

Beyond our expectations, the saccade frequency was less related with the performance of a novice pilot. The saccades, including frequency, distance and accuracy, are important indexes of acquiring and processing the information. The eye-scanning pattern is different between expert and novice [14, 15]. In our study, all the subjects were novice and only trained 8 times before the formal test. They might not have efficient saccade or link the eye-scanning activities to their performance in the simulated flight.

In the selected flight phase from cutting off auto-pilot and auto-throttle, subjects performed more tasks when adjusting the flaps or controlling landing gear, which made novice pilots take on more workload. Subjects should pay more effort and attention to maintain performance. Therefore, the correlation between the performance and multiple physiological parameters were more stronger compared with that at the Leisure time.

In conclusion , we studied the relation between the performance and multiple physiologies of novice pilots. The results in the study revealed that most physiological parameters (cardiovascular and ocular information) in a flight were relevant to the pilot's performance. Especially during tasks, the correlation was much stronger compared with that at the other time. On this basis, we will establish the model to predict the pilot's performance through monitoring the physiological parameters.

Acknowledgement. This research is supported by The National Basic Research Program of China (973 Program, 2010CB734103), China Postdoctoral Science Foundation (2013M540363) and The National Natural Science Foundation of China (61305141).

References

1. O'donnell, R., Eggemeier, F.T.: Workload assessment methodology. Measurement Technique 42, 5 (1986)
2. Rubio, S., Diaz, E., Martin, J., Puente, J.: Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. Applied Psychology 53, 61–86 (2004)
3. Lean, Y., Shan, F.: Brief review on physiological and biochemical evaluations of human mental workload. Human Factors and Ergonomics in Manufacturing & Service Industries 22, 177–187
4. Farmer, E., Brownson, A.: Review of workload measurement, analysis and interpretation methods. European Organisation for the Safety of Air Navigation 33 (2003)
5. Fowler, B.: P300 as a measure of workload during a simulated aircraft landing task. Human Factors: The Journal of the Human Factors and Ergonomics Society 36, 670–683 (1994)

6. Wierwille, W.W., Connor, S.A.: Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. Human Factors: The Journal of the Human Factors and Ergonomics Society 25, 1–16 (1983)
7. Wilson, G., Fullenkamp, P.: Psychophysiological assessment of pilot and weapon system operator workload. Stress and error in aviation (A 92-13015 02-53). Aldershot, England and Brookfield, VT, Avebury Technical, pp. 27–34 (1991)
8. Jorna, P.: Human Machine interfaces for ATM: objective and subjective measurements on human interactions with future Flight deck and Air Traffic Control systems. In: FAA/Eurocontrol ATM R&D Seminar, Paris, France (1997)
9. Miyake, S.: Multivariate workload evaluation combining physiological and subjective measures. International Journal of Psychophysiology 40, 233–238 (2001)
10. Sonderegger, A., Sauer, J.: The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. Ergonomics 52, 1350–1361 (2009)
11. Just, M.A., Carpenter, P.A.: Eye fixations and cognitive processes. Cognitive Psychology 8, 441–480 (1976)
12. Robert, G., Hockey, J.: Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. Biological Psychology 45, 73–93 (1997)
13. Yeo, G.B., Neal, A.: A multilevel analysis of effort, practice, and performance: effects; of ability, conscientiousness, and goal orientation. Journal of Applied Psychology 89, 231 (2004)
14. Law, B., Atkins, M.S., Kirkpatrick, A.E., Lomax, A.J.: Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In: Proceedings of the 2004 Symposium on Eye Tracking Research & Applications, pp. 41–48. ACM (2004)
15. Beck, M.R., Trenchard, M., van Lamsweerde, A., Goldstein, R.R., Lohrenz, M.: Searching in clutter: Visual attention strategies of expert pilots. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 56, pp. 1411–1415. Sage Publications (2012)

Proactive Safety Performance for Aviation Operations

Nick McDonald[1], S. Corrigan[1], P. Ulfvengren[2], and D. Baranzini[1]

[1] Centre for Innovative Human Systems, School of Psychology,
Trinity College, University of Dublin, Ireland
{pnmcdonld,siobhan.corrigan}@tcd.ie, d.baranzini@not.available
[2]Dept of Industrial Economics and Management, INDEK,
KTH Royal Institute of Technology, Stockholm
pernilla.ulfvengren@indek.kth.se

Abstract. The EU Vision 2020 sets a goal of reducing the air travel accident rate by 80%. Achieving this vision requires innovation and a different approach. PROSPERO (Proactive Safety Performance for Operations) is an EU FP7 project that will provide an advanced systemic methodology for managing the improvement process to help achieve that goal, as well as delivering a way of measuring progress. The overall objectives of PROSPERO are to; (i) Provide a proactive anticipation of complex system risks; (ii) Ensure more effective management of and enhanced learning from situations where risks cannot be designed out of the operation and (iii) Achieve substantial improvement in the elimination of and recovery from human error. This paper reports on the overall PROSPERO concept and high level system requirements as they emerged from the first research phase that focused on identifying industry needs.

Keywords: Risk & Performance Management, Safety Management Systems, PROSPERO.

1 Introduction

The next generation Air Transport System (ATS) requires systemic, proactive, performance-based safety management that is fully integrated into seamless ATS operational management, capable of delivering measurable performance improvement. While this is the aspiration of the current generation of Safety Management Systems (SMS) regulation, available processes, methods and tools are not adequate to realize this. The ATS has to be able to anticipate and manage complex system interactions (where each element on its own may seem acceptable) before they are manifest in operational emergencies and use operational experience more effectively as a preventive resource. Recent aviation accidents have demonstrated that this is a significant weakness of the aviation system.

'…there is not yet a universally accepted risk assessment methodology in common use across the European Union for all the aviation domains which would enable a standardised approach and better priority setting to tackle those risks that pose the greatest threat to safety. This shortcoming will have to be overcome.' (European Commission, 2011)

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 351–362, 2014.

The ICAO regulation (ICAO, 2012), mandates states to implement legislation requiring aviation organizations to have Safety Management Systems and for states to have State Safety Programs. The European Commission (2011) defines a safety management system as 'a pro-active system that identifies the hazards to the activity, assesses the risks those hazards present, and takes action to reduce those risks to an acceptable level. It then checks to confirm the effectiveness of the actions. The system works continuously to ensure any new hazards or risks are rapidly identified and that mitigation actions are suitable and where found ineffective are revised. This new approach to safety is preventive, proactive rather than reactive, it aspires to be performance driven, systemic, risk based and able to deliver verifiable improvement.

When Air France 447 took off from Rio on June 1 2009 everything was apparently normal: the aircraft was fine – the pitot tubes were an unknown and acceptable defect. The crew, aircraft, route and weather were ok – nothing unusual. Nothing changed. Yet these elements in combination created a situation that ultimately the crew could not manage. There are two ways of looking at this disaster. Hindsight: why did the crew apparently not have basic airmanship skills? The interim investigation report has recommendations about crew training (BEA, 2011). This is the classic reactive approach – address the issues arising from the most recent serious safety event. This is essential. More challenging is to ask: What could have been done before that flight to minimize the possible risks associated with it? The risks were built into the operational situation before take-off. Could routine measures in advance not just prevent this accident happening again but provide a more general preventive shield against a wide range of system accidents?

Could the implementation of the ICAO regulation (as European Directive and national legislation) prevent the type of accident that befell AF447? It is, of course, impossible to be certain, but, if one examines current methodologies and processes for managing safety, it is hard not to conclude that prevention would be unlikely given the current state of the art of safety management. This is for the following reasons:

- Risk assessment methodologies are largely based on expert judgment unsupported by extensive data and hence find it difficult to encompass complex system interactions.
- There is no integrated Air Transport System risk metric that allows risks of different types and sources to be assessed with reference to each other, singly or in combination.
- The active management of risk in planning and management of operations is not well supported
- The anticipation and preparation for potential emergencies is not integrated into normal everyday operational planning
- Because there is no system-wide risk metric it is not possible for system improvements to be evaluated against some projected risk reduction target.
- There is no standard for safety performance that a regulator can use to audit, evaluate or require and operator to improve its safety system.

The PROSPERO project seeks to address these fundamental defects in safety management in order to pave the way for the successful implementation of current

requirements for SMS regulation and to lay the basis for the next generation of Safety Management Systems,

The objectives of this paper are to present the overall PROSPERO concept and system requirements as they have emerged from the first phase of the research.

2 Prospero Objectives

The objective of the PROSPERO project is to assemble and analyse, from operational performance data, an integrated model of risk incorporating the key sources of risk: aircraft and ATM technologies, staff and crew, routes, weather. This risk information will enhance the normal supply of information into the planning, dispatch and operational management of all the components of the ATS, so that potential risks are designed out in planning or actively managed in operations, and potential emergencies are anticipated and prepared for. Feedback from the active management of anticipated risks will stimulate learning. An improvement loop will generate systemic changes to reduce risk. This risk management system will be initiated at organizational level in airline, airport and air traffic control organization and then integrated in a common ATS framework. This integrated safety performance system will be designed to be scalable up to European level and applicable to the Single European Sky concept.

2.1 Risk Assessment

Current approaches to risk assessment tend to be based on either expert judgment or extensive data-mining, but rarely both. The probability of an outcome of a certain severity is core to the conceptual definition of risk. Yet the practice of risk assessment nearly always comes down to an expert judgment, of one or more experts (see, for example, Luxhøj, 2001; Rantilla and Budescu, 1999; Ayyub, 2001). Most often this is due to absence of accurate and timely data. Bow-tie analysis, which is built around expert judgment, is at the heart of the ARMS (Airline Risk Management Solutions) methodology (Nisula, 2008). Fault Trees, Failure Mode and Effect Analysis, Human Reliability methods (e.g., THERP, TESEO, HEART, ATHEANA), Functional Analysis all rely on expert judgment. Its major limitations relate to the reliability of judgments; hence the concern with combining the judgment of different experts. The judgments of experts are not sufficiently rigorous or reliable enough to assess complex combinations of factors. PROSPERO will support the development and integration of existing databases to build a data source to support effective data-mining. This data mining will be complemented by the capacity to model and analyze the operational processes of the air transport system.

2.2 Process Approach in Human Factors

The whole idea with process control is to identify undesired outcome, produced by the process, by measuring its output and improve the process input until the desired outcome is achieved. The first step in gaining control over an organization is to know

and understand the basic process for production (Juran, 1988; Deming, 1986). In aviation the production is delivered by a complex socio-technical system. If safety is to be maintained, or even improved, relevant mechanisms influencing these processes need to be identified by a model. Improvements may be accomplished only by changing the processes' functionality. Traditional human factors and safety research does not provide any method to do a functional process analysis of socio-technical systems in order to identify relevant indicators for safety and to improve safety performance resulting in actual change in operations. A functional process model is essential for analyzing and identifying influencing mechanisms of the operations in which risk needs to be controlled. With such a model it is possible to link in-depth investigation and analysis of particular events with the extensive analysis of data from normal operations. This will deliver a composite assessment of risk.

In a series of EU-projects a method and tool called SCOPE has been developed (Leva et al. 2011, Bunderath et al., 2008; Morrison, 2009). At a concept level it has been used for adding core functionalities in Safety Management Systems in both airline and airports. Some of the functionalities have been system description, hazard and indicator identification, support for process analysis to identify needs for improvements and to simulate and anticipate effects of system changes. In PROSPERO SCOPEs value for analyzing integrate and systemic risk will be further developed and validated.

2.3 System Risk

Despite the interdependencies between all the components of the Air Transport System there is no integrated system risk concept. Even within an airline between flight operations and maintenance risk means different things to different parts of an organization each of which have different baselines and priorities. For example, deferred defects may be an acute immediate risk for a maintenance organization, but not high on the priorities for flight operations. Therefore it is important to establish a common framework of safety performance because it is these mutual interdependencies that ultimately determine system risk. However a challenge in achieving this is the lack of one institutional owner of a system risk concept.

PROSPERO will facilitate the development of an integrated risk framework that establishes sufficient commonality between safety performance indicators to support an appropriately integrated analysis of risk. PROSPERO will integrate Human Factors in a systemic analysis of risk, based on the innovative SCOPE methodology (Leva et al. 2011), which analyses how human, social, information and other factors combine to influence system performance. It will develop an integrated concept of risk that equally reflects human behavior in the mission of the crew, maintenance personnel, and ATC and ground operations. The PROSPERO methodology will provide for a proactive anticipation of complex system risks that have the potential to give rise to abnormal situations and crises.

2.4 Operational Management of Risk

Risk analysis and assessment should be part of a risk management process that concludes with an evaluation of risk reduction following the implementation of measures to mitigate and control the risk; or where it is not possible to mitigate the risk (through design, process change, planning, etc.) there is active and explicit management of operational risks in real time by crew during operations. Hence the quality of the management of risk is dependent on the quality of the initial assessment of the risk itself. Unfortunately current operationally focused risk management methodologies (for example, Threat and Error Management TEM) are not integrated with an effective risk assessment methodology. An Intelligent Flight Plan developed in the HILAS Project (from an Iberia use of TEM) is a smart concept for improving operational management of risk, incorporating an operational risk assessment in the normal flight preparation process rather than having it as an extra task with more effort involved (Cahill, 2009). This can be developed further by incorporating a comprehensive, authoritative and up-to-date account of operational risk.

PROSPERO will take this as starting point and build a much stronger integrated concept, capturing risk information produced by airlines, airports and ANSPs and embedding it in the normal flow of operational data. This is made possible by the increasing digitalization of information for flight planning and management. For example, the EFB ('Electronic Flight Bag') creates an interface for flight crew and maintenance for a wide range of technical information about the aircraft and enables the technical log to be managed in digital form. Increasingly route and weather and crew rostering information is being fed in digital form into the flight planning information systems to provide an automated service for planning, dispatch and crew planning and briefing. PROSPERO will use this infrastructure to embed targeted risk information so that not only crew, but planners and human resource managers can begin to identify more sharply potential discrepancies between the resources supplied (crew training records, for example) and the risks identified in a particular operational configuration.

2.5 Anticipation of Emergencies

The accident involving AF447 is just one example (amongst many) of a lack of preparedness for a potential emergency becoming manifest in inappropriate control actions inadvertently escalating the situation. Mental preparedness for an emergency is critical in ensuring appropriate response. Airports have a statutory requirement for periodic major emergency exercises, and simulation training for flight crew includes special training of non-normal processes and coping with particular types operational emergencies. Nevertheless there seems to be a gap in the routine priming of emergency preparedness. In PROSPERO, the provision of smart up-to-date and targeted risk information about an operation being planned will provide the opportunity not only to plan and prepare for how to manage such threats in a normal way, but also to rehearse potential emergency scenarios that are relevant to that particular operation.

2.6 Improvements in Managing Risks

Improvement processes are weak in aviation as in other industries. Research in another EU funded project AMPOS showed that each stage in the improvement cycle is more difficult than the last. It is almost impossible to get evidence of evaluation of implementation of recommendations. It is important to find ways to maintain awareness of potential risks, at the operational level, including many hidden risks from time delays or deficiencies in the technical safety process. However there is no clear methodology for ordering and comparing the risks arising from different system defects or process deficiencies and hence for prioritizing improvements or for maintaining a high level of risk awareness at operational level. The PROSPERO project will take significant steps towards developing such a common risk framework that can maintain active attention to potentially significant risks at operational level; this will in turn provide a basis for prioritizing initiatives and projecting a reduction of operational risk following implementation of change. This will also enable the tracking of risk reduction during the processes of technical and organizational change.

2.7 Regulation of Safety

Of the air transport system only the air traffic management aspects are subject to performance regulation at European level (the Single European Sky). The Performance Review policy for safety of the SES is currently based mainly on one dimension - implementing Safety Management Systems, without the support of independent safety performance indicators. Without such independent criteria neither the regulator nor the regulated agencies can show measurable improvement in safety. Furthermore the performance of the ATM system is itself dependent on the performance of airlines and airports. Hence the development of a meaningful performance concept at the core of aviation regulation really requires an integrated whole system approach.

PROSPERO will provide a prototypical whole system approach to aviation performance that can be scaled up to European level. It will provide the methodologies for measuring and analyzing safety performance and assessing risk, which are essential for measuring improvements (or otherwise) in safety performance for the air transport system.

2.8 Safety Management and System Change

The aviation system is changing rapidly. New business models are transforming operational norms. There are major technological initiatives such as SESAR which are bringing new processes and operational concepts. Do current safety management capabilities meet this challenge? It is possible to trace an evolution of safety management as it attempts to address these challenges.

In Stage 1, Classic safety management, safety acts as a brake on change. Static safety standards provide a fixed reference point against which the system is evaluated. “Safety margins” maintain an uneasy balance between the opposing forces of safety and cost. Safety is managed as an independent system within the organization with

little leverage over operational change. Stage 1 is typical of the JAA Regulations from the late 1980s and 1990s.

In Stage 2, safety management has to provide assurance in a time of change. Change erodes safety margins, for example in flight time limitations. The boundaries between what is safe and what is not safe are no longer clear. Active risk management is necessary to monitor system safety, for example in the development of fatigue risk management. Safety failure is a major corporate threat, because the leading business model is about reduced margins & lean processes. Safety Management System models are built on aspirations to be proactive and systemic, with limited guidelines on implementation. A good exemplar of this model is the work done by easyJet, both in the HILAS project and outside of it, in developing a corporate strategic risk management approach and implementing this in a much more dynamic fatigue risk management process.

PROSPERO is part of the evolution of Stage 3 in which safety management is a partner in change. Rigorous operational safety analysis delivers robust processes that give high reliability from all points of view. Safety is part of an integrated performance management concept with common performance indicators, integrated risk management and a common change program. All of this manages the transition from present to future – projection of future process is based on full modeling and analysis of all implications and risks. The approach integrates culture and system management showing how to change the way the system works in order to influence culture. In this model a strong performance concept delivers an independent operational criterion of adequacy. This gives confidence that change can be managed against a rigorous criterion of operational effectiveness. When SESAR delivers its new information systems and new operational concepts into the air transport system, it will be necessary to have a robust performance management framework to monitor its safety effectiveness. This is beyond the scope of SESAR, but it is necessary in order to prepare for the full operational implementation of SESAR. Safety management in Stage 3 is highly complementary to the EOCVM (European Operational Concept Validation Methodology), the standard that governs the validation of SESAR, in that PROSPERO seeks to provide a methodology for operational evaluation at the implementation phase, thus offering the potential to link safety assessment at the design stage with safety assessment at the operational stage. This could be critical to demonstrating the success of the next stages of SESAR.

3 Methodology

The first phase of the project focused on identifying the industry needs, further validating, defining and developing the overall operational concept and guiding the development of the functional specification of the system. PROSPERO adopts a needs-based approach that can be validated against end-user acceptability and stakeholder assessment of domain suitability. However PROSPERO also has a research-based approach for industry; it builds on a series of previous research projects which have developed partial solutions to fundamental problems, identifying needs, gaps

and theoretical challenges which are critical to the development of next generation systems.

Investigations and field studies (e.g., interviews, document analysis, observations, work-shops etc.) have been conducted focusing on the risk information systems and organizational learning processes for managing safety performance described in the PROSPERO model.

- A number of bilateral partner meetings over the first phase of the project
- Two workshops with several service providers forming a local system was conducted, one in Athens and one in Rome.
- Ten interviews were performed with industry or regulatory representatives
- Survey was launched for Star Alliance partners (six replies so far)
- A common risk classification study was performed with four industrial partners
- Several approaches to risk assessment were studied and documented in literature reviews.
- Several PROSPERO partners have been involved in related research projects prior to and leading to PROSPERO (e.g., ones HILAS (EU FP6) and MASCA (EU FP7). Reports and part-deliverables have been shared and previous research was presented at a work-shop so as to give all a chance to understand the concepts that PROSPERO builds on.

4 Findings

This section provides an overview of the key findings from the first phase field research conducted at organizational level, at regulatory level and at system level (and combinations of the above). Many of the same challenges recur at the various system levels. The required flow of data and information follows the people and goes across system boundaries. The risk dependencies follow operations that cross system boundaries. Hence integrating data, people and risk are a common theme. PROSPERO aims at improving existing solutions as well as inventing new solutions for producing, distributing and using Risk information for both system change and in operations risk management.

4.1 Concepts and Needs

Risk information production and operational risk management

The following sections explain the production and use of risk information in operations as illustrated in figure 1 below which presents an overview of the overall PROSPERO concept as it evolved from phase 1 of the research.

The premise that ‘the risks are built into the operation before the flight takes off’ sets up a requirement for the availability of the most up to date risk information specific to that operation (e.g. that flight). Flight planning and dispatch need to design out or mitigate as many of the known risks to that operation as possible. Subsequently, an ‘intelligent flight plan’ can support active ‘threat and error management’ briefing for

crew that enables anticipating and actively managing those risks throughout the operation.

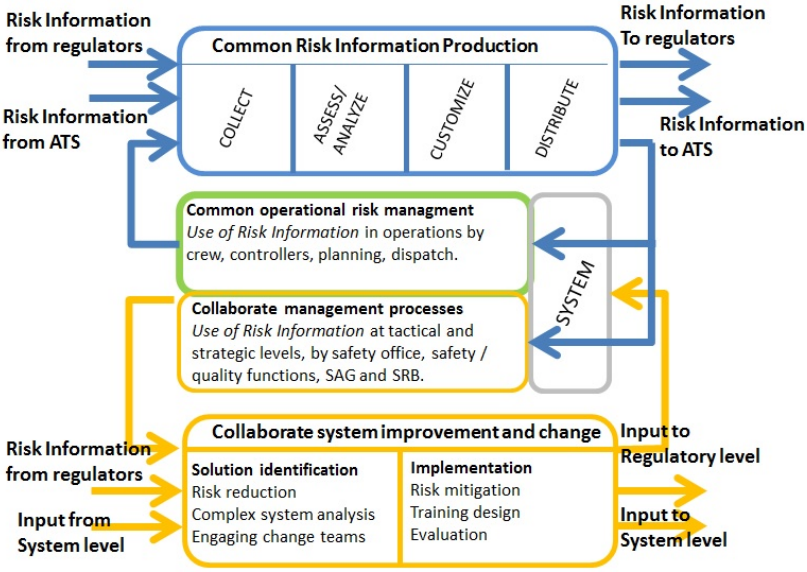


Fig. 1. Overview of PROSPERO Concept

For that to happen, risk information needs to be configured, or customized, specifically for that operation, and preferably in a way in which potential interactions between diverse types of risk can be understood and anticipated. In the diagram this is represented as the "customization" of system risk information according to the specific requirements of each individual flight plan. Aviation information service providers may have their own sources of specific risk information, related to particular technologies or information systems that need to be integrated into this targeted risk picture. It is necessary to attribute the risk to specific antecedents ('causes, contributory factors') that need to be taken into account and embedded into the normal flow of flight planning information. These relate to a range of domains: a/c technologies, crew, route, weather etc. Thus risk management becomes a normal part of planning and operational management and not an extra task.

This requires that there is a sufficiently sophisticated and detailed "Operational system risk model" that can support the disaggregation of risk information into specific antecedents, and the recombination of that information to target specific operations. To achieve this, the system, risk and data analysis needs to be statistically and analytically sophisticated to ensure that a wide range of factors and their possible interactions are covered. This will necessitate the integration of a wide range of data sources – both outcome measures of various types as well as inputs to the system. Hence inputs to the flight plan are potential antecedents to risk in the operation and thus a critical source of data. This tries to ensure that 'leading indicators' are as upstream as possible. All of this depends upon comprehensive data collection, acquisition and

reporting, including the requirement to report on how risks, identified in the 'intelligent flight plan' were managed in each operation.

While there is a commonly recognized requirement for a quantitative approach to risk assessment, this poses significant challenges from a data science perspective, concerning, for example, the availability, quantity, quality and compatibility of data to support such statistical analysis.

4.2 System Improvement and Change Process

Risk information production and use by management for the system improvement and change process are illustrated in the figure 1 above. The ultimate goal is to be able to demonstrate the impact of a new system, or an operational change, on operational outcomes. This requires a clear set of risk dimensions that have driven the whole design, development and implementation process, against which the new design or change implementation is monitored and evaluated. At the beginning of the implementation phase it is necessary to be able to project (and then manage) the risks of implementation. These come from the combination of the functional characteristics of new technologies, operational processes, social systems, information and learning systems. The process of design, development and implementation needs to conclude with a validation of the implemented changes and improvements against those system risk requirements.

The risk analysis needs to be sufficiently data rich and statistically and analytically sophisticated to encompass those complex human, social and technical system interactions, so that it can support the modeling and projection of future risk scenarios in the 'to-be' system. This, in turn depends on the integration of a wide range of data sources both from the operation, as well as those specifically related to the technology, so as to ensure that all risk assessments in the process are ultimately accountable to a credible projection of operational impact. For this to work at an ATS level, there need to be common processes, methodologies, taxonomies, etc., for assessing, analyzing and classifying risk. At the management level, teams need to collaborate with stakeholders who are relevant to what needs to change and who can have an impact on that change. The regulator has an important role in producing risk information that supports the system and organizational level in improving their operational processes and safety performance. This role needs to be enhanced.

The two themes of industrial needs that were strongly articulated in the field research are quantitative risk assessment and the common integrated risk analysis. The field research did not show as strong evidence for industrial needs for the other three areas: operational management of risk, preparation for emergencies, risk-managing design and change, and regulating risk. These needs in these areas were clearly identified in previous research. The research themes of Quantitative risk assessment and the Common integrated risk analysis are both essential parts of the Risk Information Production function. In the revised model of the PROSPERO concept, risk information production includes the steps of: collecting data, risk assessment and data analysis, customization for intended use and distribution of the risk information to the user(s). Of these functions the theoretically most challenging area is the risk assessment.

When this is resolved it will enable a "pull" of identified relevant risk data from various sources, system levels and stakeholders.

5 Conclusion

The first phase of the research evolved the overall concept and highlighted the system requirements as depicted in figure 1 above. The research and development needs of PROSPERO so far has involved a complex combination of methodologies, information systems and software, organizational processes, and social and business relations between stakeholders. Each of these requires a specific research focus, but not in isolation - rather, as a multi-layered systems integration project. The next phase of the research is to specify the functional specification of the actual PROSPERO system. This will also involve understanding how the identified industry needs should be expressed as a set of requirements against which PROSPERO can clearly demonstrate the feasibility (in terms of technical criteria, rules and procedures, organizational capabilities, etc.) of a data-driven framework with common methodologies. All of which is required to understand risk in air transport as a system, and to show how to use that knowledge to mitigate risk in everyday operations and to radically transform the risk through future system design.

References

1. Ayyub, B.M.: Risk Analysis in Engineering and Economics. Chapman & Hall/CRC (2001)
2. BEA - Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, Final Report to AirFrance AF447 Flight (2011)
3. Bunderath, M., McDonald, N., Grommes, P., Morrison, R.: The operational Impact to the Maintainer. In: Proceedings of IET Conference, London (2008)
4. Cahill, J., Ulfvengren, P., Gaynor, D.: Performance Management. HILAS deliverable. Project Number 516181.EC - 6th FP (2009)
5. Deming, W.E.: Out of the Crisis. MIT Press (1986)
6. European Commission (2011): Setting up an Aviation Safety Management System for Europe. Communication from the Commission to the Council and the Eur. Parliament, Brussels, October 25, 2011 COM (2011) 670 final
7. HILAS, Human Integration Into the Lifecycle of Aviation Systems. Priority No: 1.4 - Aeronautics and Space Research Area 3. Priority Title: Improving Aircraft Safety and Security, IP8. Proposal/Contract No: 516181. Brussels: EC (2005-2009)
8. International Civil Aviation Organisation, Safety management manual (SMM), 3rd ed., Montreal, Canada: Doc 9859 (2012)
9. International Civil Aviation Organisation, Safety management manual (SMM), 2nd ed., Montreal, Canada: Doc 9859 (2008)
10. International Civil Aviation Organisation, Safety management manual (SMM), 3rd ed., Montreal, Canada: Doc 9859 (2012)
11. Juran, J.M.: Juran's Quality Control Handbook, 4th edn. McGraw-Hill Companies (1988)
12. Leva, C., Ulfvengren, P., Corrigan, Z.R., Baranzini, D., McDonald, N., Licata, V.: Deliverable: D 2.1 Review of requirements for change Deliverable to EC as part of MASCA project FP 7-AAT-2010-4.3-4.: Grant agreement, vol. (266423) (2011)

13. Luxhøj, J.T., Maurino, M.: An Aviation System Risk Model (ASRM) Case Study: Air Ontario 1363. The Rutgers Scholar 1 (2001), http://rutgersscholar.rutgers.edu
14. Morrison, R.: Operational Process Modell / Knowledge Space Model HILAS deliverable and draft manuscript to book. Project Number 516181. Funded by European Commission - 6th FP (2009)
15. Nisula, J.: Operational Risk Management, Work by the ARMS WG, Delivery report, ECAST 2010 (December 2008)
16. Rantilla, A.K., Budescu, D.V.: Aggregation of Advisor Opinions. In: Proceedings of the 32nd Annual Hawaii International Conference on Systems Science (1999)

Participatory Design of a Cooperative Exploration Mediation Tool for Human Deep Space Risk Mitigation

Donald Platt[1]. Patrick Millot[2], and Guy Andre Boy[1]

[1] Human-Centered Design Institute, Florida Institute of Technology,
Melbourne, FL 32901, USA
{dplatt,gboy}@fit.edu
[2] LAMIH CNRS, University of Valenciennes, France
patrick.millot@univ-valenciennes.fr

Abstract. This work describes the participatory design and development of a Virtual Camera (VC) system to improve astronaut and mission operations exploration efficiency and cooperation while exploring in deep space. Advanced interaction media capabilities can improve exploration efficiency and cooperation as the distribution of human space exploration roles change in deep space exploration. This capability was developed in a tablet-based application that was evaluated in the field. The VC can minimize the risk of astronauts exploring unknown reaches of the solar system with limited previous knowledge of the area under exploration. Ground-based expert knowledge can be captured and be easily assessable in the remote deep space environment with the VC. The human-centered method of development and testing is described as well as results.

Keywords: Situation Awareness (SA);Augmented Reality; Human-Computer Interaction (HCI); Tablet Computing; Usability Testing; Space Exploration.

1 Introduction

The Virtual Camera (VC) is not a specific interface or a single platform. It is a database exploration concept providing the ability to interact with the database with multiple interaction methods. The VC concept is an agent to assist in enhancing future space exploration. With the orchestra organizational model [1] all actors are acting with the same sheet of music or are on the same page. As organizations progress in the 21st century they are no longer well suited for a hierarchical structure. Roles continue to change and be redefined, much like how the roles and emphasis changes in a complex piece of music. Human deep space exploration is facing this challenge. The VC is designed to help in this role re-assignment process. Goals, such as areas to be explored, can be shared among team members so that everyone is aware of what should be explored, what team members are currently exploring and also what areas should be avoided.

There is a continuous human-human interaction in space exploration that will be limited due to communication delays caused by distance in deep space exploration. The

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 363–374, 2014.

control model is switching from one of supervision from the Earth-based controllers to one of mediation between the ground and the astronauts, see Figure 1. Risk will be increased as astronauts roles change to one of more autonomous decision making as opposed to ground-mediated decision making. Risks such as decision making based upon incomplete information or a lack of domain knowledge are very real for astronauts exploring in deep space. Real-time decision making will be required of astronauts without real-time support from operators back in mission control due to the communications delays in deep space. Tools will be required that support the astronaut's knowledge-based reasoning and abduction to collaborate with human judgment.

An important aspect of the VC is the idea of improved team or multi-agent situation awareness for exploration. This takes into account the goals and requirements of each individual user. Rather than being concerned that everyone involved understands the same thing, the VC is designed and used from the perspective of understanding how each user sees the exploration goal from their own unique perspective. This is much like the musicians in an orchestra. They all have their own instruments and parts of the music, but they need to play together to make art coordinated by a conductor. Human deep space missions will be characterized by remote distributed operations. There will be a need to make decisions based upon the collection and analysis of raw data to provide predictive information. This information needs to be presented to crews in a way that enhances situation awareness.

The Virtual Camera (VC) is an interactive 3-D tablet device that allows astronauts to capture expert knowledge about the areas they will explore [2]. The VC integrates and manages mission data and areas of interest for both exploration and safety. This information will include system health and status, caution and warning, safety, traverse execution and mission timeline parameters. This paper explores the VC and these changing roles for deep space human exploration and the new human-machine cooperation needed for deep space exploration and how this system can mitigate risk in human deep space exploration. Testing and evaluation explore how the VC influences interaction and cooperation between actors involved in human spaceflight.

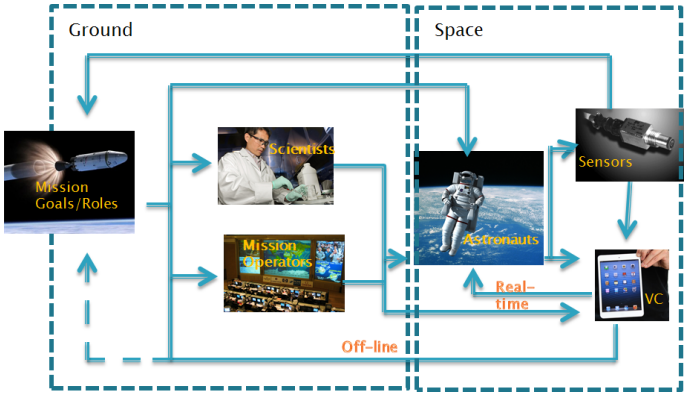


Fig. 1. The model of the virtual camera interaction with a space segment, used by astronauts and a ground segment used by mission operators and scientists

2 Motivation

Interactive cockpits, context-sensitive systems and remote agent knowledge representation are being developed for a number of domains. Modern users of computer technology are very familiar with and aware of tablet-based interaction technology. This technology should also be applied to the next generation of human space exploration systems. The VC demonstrates these possibilities in a system usable for deep space human exploration.

Astronauts were familiar with relatively simple analog and CRT displays in the era of the space shuttle. This is now changing with new "interactive" cockpits. Older astronauts will be less familiar with the concept of interactive cockpits but younger ones will most likely have used them for most of their flying careers. Tablet and other advanced interaction media systems are now commonplace in a number of domains including home usage and stakeholders for the VC will have exposure to this type of interaction making it a natural for future space exploration systems.

Mission operations personnel will be interested in using the VC for mission planning purposes as well as training. For the actual mission the large distances involved between Earth and the body being explored means that communications delays will make it impossible to provide real-time feedback to astronauts. This requires the VC to capture the expertise of the mission operations personnel on-board. The goal is to make the VC a remote assistant/agent for the mission operations personnel.

Scientists will also use the VC for planning future traverses and to look for new discoveries on the body being explored. Their experience level with the VC tool will typically be less than the astronauts and mission operations personnel. They will want to denote and annotate areas of scientific and exploration interest to them. This information can be archived late retrieved by other VC users. The virtual camera can provide a third-person perspective to assist in the navigation process for rover and other exploration vehicles. However, it is more than just a "back-up" camera that is now provided on many terrestrial automobiles.

The VC was implemented on a portable tablet computer offers many unique advantages including the ability to bring the device around the cockpit to any required vantage point to explore the remote planetary surface. For instance, a crew member using a tablet-based VC can carry it easily from one window to the next as the vehicle maneuvers around areas of interest. It can also carry electronic procedures that can be pulled-up and displayed depending upon the situation. The VC can also be carried by outside explorers who are walking across the surface of the planetary body. Most of the testing presented in this work was completed in this scenario, with users carrying the device while navigating and exploring on foot. Other possible uses and configurations include using the device in a planning scenario in conjunction with large format and three-dimensional displays on earth or at a habitation base. A portable, tablet device allows improved interaction, collaboration and sharing of new knowledge to improve efficiency of exploration. In this case, efficiency is measured by the ability to quickly identify areas of interest to explore and then to explore more areas in a given period of time. Greater situational awareness leads to higher levels of efficiency because additional areas can be explored, i.e. more knowledge can be gained (and

retained) on a single expedition, thereby reducing the number of expeditions required. It provides one integrated tool for navigation and exploration assistance as well as annotation.

3 Related Work

Endsley [3] broke situation awareness down into three elements, perception, comprehension, projection. The VC contributes to improving exploration situation awareness on all three levels, see Figure 2. Perception is improved by offering multiple views and perspectives of the scene around the user. This could be correlated with vehicle position data to show a real-time situation or it can be used as a training element showing where a vehicle may be at a certain time during a traverse.

Using the VC for deep space exploration, comprehension is improved by allowing the user to understand more about their environment than what is possible just by looking out of the window of the vehicle. With an annotation capability knowledge from domain experts can also be incorporated into the database. For instance, mission operations people can annotate areas of particular safety or science interest during training and these notes and other information can then be pulled up when the astronaut is actually at that location on the remote planet.

Future deep space human exploration will require a transfer of function allocation from Earth-based mission control to on-board the spacecraft. Communications delays due to distance will require crews to be more autonomous and not rely on mission control for real-time advice and control. For example, the time delay in communications to the moon will take 6 seconds round trip, and up to 44 minutes for Mars. For missions lasting perhaps more than a year training conducted on the ground may be forgotten. Skill-retention issues will be a concern. On long-duration space station missions today, astronauts often find they are pulling information from memory in what becomes a subconscious effort. Training can take place months before the actual space-based operation takes place. This capability to pull from memory the skills needed to complete the mission may be diminished when practice is no longer possible during life-critical deep space mission. On-board support and refresher training will be required. The VC allows an astronaut to move through a traverse or exploration sequence ahead of time. Tools are currently being tested to capture mission control knowledge and expertise on-board a deep space exploration vehicle.

Projection will be improved by allowing traverses to be rehearsed ahead of time. The best route can be determined and planned. Safety and exploration performance considerations can be taken into account. Possible future actions can be modelled and best choices made. Millot and Hoc [4] defined two concepts in which cooperation between humans and machines takes place, the first is know-how and the second is know-how-to cooperate. Know-how is the abilities or information of a single agent. The latter involves agent-to-agent interaction. In the case of the VC the know-how is being transferred from ground personnel to astronauts through a mediation agent, the VC. This provides a common work space for the interaction.

Dynamic allocation of tasks will be required in human deep space exploration. Mission control, which may then be called mission support in this case, may be able to make some mission decisions but oftentimes the astronauts will be required to do so on their own. Pacaux-Lemoine and,Debernard [5] broke down human-human agent cooperation during possible decision interference in three ways, negotiation, acceptation and imposition. For the VC interaction, negotiation involves improving the frame of reference of ground personnel and astronauts and exploration goals. Certain mission goals may take priority at different times and the VC can help this prioritization. Acceptation will involve agreement on the next exploration course of action. Finally, imposition will require defining next course of action without negotiation in some cases when for instance communication blackouts or emergencies cause a decision to have to be made without the ability to interact at all with Earth.

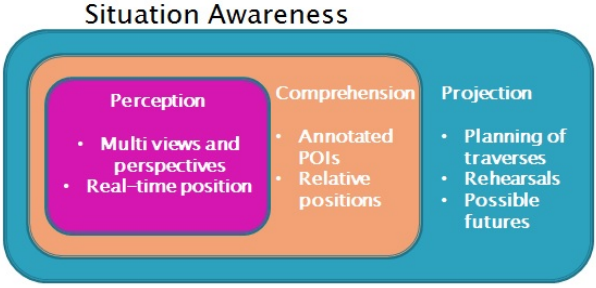


Fig. 2. The VC allows improved exploration situation awareness on the levels of perception, comprehension and projection

Stanton has suggested that teams cannot truly share situation awareness [6]. In a collaborative environment such deep space exploration we need to be concerned with distributed cognition where appropriate information, held by individuals but captured by devices represents s dynamically changing environment. The VC must transfer this appropriate information to appropriate team members at the appropriate time.

Collaboration can be mediated through computer systems in what is known as computer supported collaborative work (CSCW). Collaboration can involve a few individuals or a team. CSCW is concerned with how technology can effect collaboration [7]. Grudin and Poltrock also looked at human behaviors that contribute to collaboration and broke them down into three main categories: communication, sharing of information and coordination. Collaboration can be real-time or asynchronous. Systems that enable or enhance collaboration should enhance these behaviors in real-time or asynchronously. Modern social media technologies such as Twitter and Facebook combine communications and information sharing capabilities. Social units are also important, such as small groups, teams, projects, organizations and communities.

The VC encourages user interaction within its database. According to Grudin and Poltrock, determining how to route, organize, and present contextual information to facilitate collaboration is a pressing challenge. They also suggested that qualitative

field research is an area with unlimited potential. New technology deployment always provides phenomena of interest. Qualitative research can find patterns that are then followed up with qualitative research to discover what the patterns mean. Collaboration and collaborative exploration bring to mind teams and the concept of team situation awareness. There has been considerable debate about the usefulness of team situation awareness and even how to measure it [6]. Team members may have different immediate goals and certainly different roles within the team environment. Therefore, it may be a detriment for all team members to share the same situation awareness [8]. In fact, the cognitive properties of a group may differ from the individuals who make up the group [9]. Even for analyzing individual's situation awareness, it is difficult to apply probes such as Endsley's SAGAT during real-world collaborative tasks [6]. Reality often involves events happening in parallel and serial models do not do well in accounting for this.

4 Participatory Design

A human-centered design approach has been used to develop the tablet-based VC. The first step was to develop low-level prototypes to capture the basic design requirements for the VC. The initial point in determining user requirements is to identify experts that represent all possible user types for the VC system. This requires some knowledge of the domain and also of the goals of the VC. A survey of potential users and stakeholders (astronauts, mission operations personnel, and scientists) for the VC system was conducted by the authors at the 2011 NASA DesertRATS analog exploration testbed. The main questions during potential user interviews involved their background, general system uses, interface type and data display parameter formats desired.

Brief scenarios and interaction diagrams as well as story-boards and prototypes have also been developed. These can be presented to users to show the basic concept of the VC and to get feedback of what would be a beneficial inter-face for a particular user and what may also not be of benefit to a user. Scenarios and use cases give users examples of how the VC can be applied. Next horizontal prototypes were developed that captured the basic interaction requirements for the VC and were used to demonstrate an operational scenario.

A complete vertical prototype was then developed with all of the interaction capability using a tablet PC device. Usability of the VC will be evaluated using nominal and off-nominal situations and simulations of actual mission operations cases. Important considerations include what is appropriate for a given task or function in a given context such as after an alarm or interruption versus normal operations.

Usage of the vertical prototype determined how unique aspects of the tablet improve or impact situation awareness, exploration efficiency and collaborate.. The completed tablet VC is a tool that is used to analyse ways to improve situation awareness as well as illustrate how interactions between collaborators and their various roles change in deep space exploration compared to low-earth orbit (LEO) operations.

Usage of the system in the field also illustrated other applications for the VC beyond space exploration as well as emergent behaviors.

5 Evaluation

Preliminary testing of the VC has been completed to identify the interaction capabilities and issues when operating in deep space Initial tests have been conducted in a terrestrial setting (a public park) with areas of interest (AOIs) defined to be of science or resource use as well as hazards. These points have been entered into the VC Google map terrain database. After lessons learned in the first set of evaluations, a second set were conducted with a team of scientists in a true science exploration setting.

For the first round of testing three teams were involved with 8 different subjects in each who all had tablet experience. The goal was not to gather a large statistical sample but rather to gather feedback and collect emergent behaviors and uses of the system. The subject team used the VC to navigate a course encountering AOIs of various point values to simulate mission importance scale. The order of AOIs was defined by a Mission Control (MC) team ahead of time in consultation with the VC subject team. A control team will also be on the same course using a non-interactive tablet-based map of the area. The control team used a feature limited version of Virtual Camera which can display static AOIs and report data to MC. The MC team had access to the same database of AOIs as the Subject and Control teams, although communication between the teams in the field and MC team involved a simulated lightspeed delay. The communication delay was done by having a phone in the possession of the observer of each of the two field teams who did NOT answer the phone. The MC team left voice mail responses. The observer waited an agreed upon time delay, then played back the voicemail. Thus half duplex, time delayed messaging was simulated. This simulated the changing environment the VC is expected to operate in. Pictures were taken by the field teams to confirm that each resource site was visited correctly.

Observations were made of the test subjects in the field. They were asked to speak aloud and give their impressions of the use of the VC or map-only interface to assist them in determining what to explore. Questionnaires were used to evaluate navigation situation awareness during VC testing and to compare it to teams in the field who had the map only for navigation. For navigation situation awareness, at two points in each 30 minute traverse both the VC and map only subjects were asked to answer a series of questions about their ability to perceive their situation, comprehend it and project into the future. The observer then determined whether the answer was appropriate or not for the given situation. This situation awareness assessment tool is referred to in this study as the Knowledge-Based Situation Assessment Tool (KB-SAT) and its characteristics are:

- Develop a series of questions to be asked of users in the field
- questions should probe at one of the three stages of situation awareness formation (perception, comprehension and projection)

- Give a brief (30 second) period for the subject to respond, using the device being evaluated (VC)
- Log whether the response was appropriate or not (for example: was the nearest object of interest identified correctly?)

Questionnaires based upon the KB-SAT situation awareness level protocol as well as the NASA TLX workload index, the SUS Likert scale and the Modified Cooper-Harper tool were given the teams.

The subject teams using the full VC interaction system found themselves relying more heavily on the VC tablet application to the point that when the wireless link to the Google database was lost the team felt somewhat confused about their location. This shows the need for data validation and a thorough review of data in the VC by system experts before and after each traverse or sortie. In the future, icons could also change color or intensity based upon confidence level of data entered into the database.

Debriefings were conducted of teams after their in-field evaluations. They were asked to rate their confidence (trust) in navigation, location finding and identifying exact position of themselves and AOIs/hazards in the field with either the map only or the VC. They would select a value from 1 to 5, with 1 being no confidence and 5 being complete confidence. Using feedback from these debriefings, it was determined that using the map alone with none of the interactive capability of the VC, the control team found they had high confidence in where in the course they were but lower confidence on what exact AOIs they were looking for, see Table 1. They did not have the ability to see the AOIs marked off at a higher zoom-in level and had no access to the annotation information placed by the simulated ground experts in the database. They found themselves surveying actual landmarks often to get their bearings. This indicates a reliance on the map for navigation but more unknowns about the actual AOIs and what the team was actually looking for. Since time is precious during space-based exploration, efficiency and precision are very important.

VC users were operating in a much more “heads-down” mode relying more heavily on the technology. The breadcrumb display was useful for finding where the team was and the direction they were headed, much more useful than the large heading arrow. The VC teams felt with a high level of confidence where they were headed and where the AOIs were. This indicates the VC is quite useful for defining, identifying and locating the AOIs.

Table 1. Confidence ratings (scale 1 to 5) for navigation and AOI identification for both VC and map only (averages and standard deviation)

	VC Navigation	VC AOI Identification	Map Only Navigation	Map Only AOI Identification
Average	4.4	4.0	3.4	3.0
Std Dev	0.79	0.58	0.55	0.71

Tests were conducted with a simulated communications delay between the field teams and mission control. This was done by having an observer communicate by text message with mission control in order to answer questions. The mission control team would then wait a predetermined amount of time and then respond. It was observed that the VC field teams would often not wait for responses to questions that they texted to mission control, especially considering the 30 minute traverse time limit. They would use the VC and their own knowledge to make decisions and begin improvising and completing the action without waiting for a response and move forward with their exploration and search for the defined areas of interest. When the team did receive the mission control text response they would use the response for confirmation of further action, relying on their own decision-making capability. The large risk associated with the time delay between mission control and astronauts in deep space exploration will require local tools that capture some of the remote knowledge and expertise. This highlights that the VC needs to enable the interface that makes the recall and interpretation of this captured knowledge as seamless as possible. Also the VC enables a closer collaboration and sharing of information between several astronaut teams who may be exploring in the field. The explorers can add information about sites of interest and scientifically important items as they are uncovered in the field. This also illustrates the changes in collaboration and cooperation for deep space missions. To expedite the exploration process astronauts far from Earth will not be able to wait for responses from Earth and require on-board decision making assistants such as the VC. They will be switching between what Hollnagel [Hollnagel, 1998] and Stanton [Stanton, et al, 2008] described as phenotype and genotype Schemas. They may revert back to training or previous knowledge if the VC is not operational.

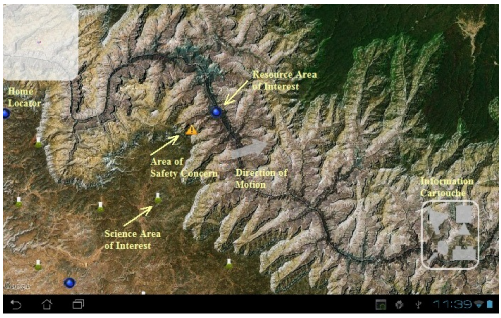


Fig. 3. The virtual camera interaction system showing the points of interest and information display

To determine the possibilities offered by the VC and to evaluate its use in an exploration environment, the VC was used at an astrobiology scientific expedition involving a science team led by Dr. Chris McKay of the NASA Ames Research Center and a group of teachers from around the state of Idaho. The evaluation was conducted as part of the Spaceward Bound summer exploration session at Craters of the

Moon national monument in Idaho sponsored by the NASA Idaho Space grant consortium. This unique park, with lava flows and volcanoes similar to those found on the moon, was used to train the Apollo era astronauts before they went to the moon.

Astrobiologists and geologists from NASA Ames Research Center, Idaho State, Idaho University and teachers mapped the lichen growth at a variety of places around the Craters of the Moon national park using the VC as an assistant. One goal was to compare the ability to conduct field science with traditional tools of a digital camera, GPS device and field notebook to using one device, the VC. Also, collaboration using the VC to map scientifically important objectives was observed. Three days of field use studies were obtained while the team mapped geologic and biological (moss, lichens) components in lava flows. Overall, the VC was evaluated for science utility, collaboration and use for education.

Evaluation was conducted by training the scientists on the use of the system and then allowing them to take it into the field to use it during their field studies, see Figure 4. A goal was to make the use of the VC as noninvasive as possible for the work they were conducting. Part of the goal of the expedition was to use a hand-held x-ray fluorescent spectrometer to map the concentration of elements in the rocks around lichen outcroppings. This device produced a read-out of the elemental concentrations. The VC was then used to capture the element compositions and these were then combined with a picture taken of the site showing where the data was collected and other scientific notes about the sample. The VC then provided a geo-located scientific record of the measurement including observations and other documentation, see Figure 5. This figure shows the record as seen in the web page database that can be called up from any computer on the internet, showing the exact location the sample was taken.

Fig. 4. A scientist using the VC to collect and annotate science data in the field

The points of interest, of course, can also be pulled up within the VC tablet interface as well. Using the VC interface points in the database can be viewed or edited. Additional information can be added as well. So, before or during a traverse explorers can pull up the information and see what points are nearby and what data is available on them. For the field tests, where it should be noted the scientists were evaluating the device on their own, they rated the VC about the same as current systems for interaction and facilitating scientific discovery. The encouraging results are

that overall the VC rates very favorably for scientific exploration for these scientists and in no evaluation point did it rate worse than current systems. Conflict resolution and finding points to explore faster were the two areas with the highest comparative rating for the VC. This is encouraging as well for a device intended to be used as a decision support aid and improvisation tool when outside expertise may not be available.

Several different types of data were gathered such as, situational awareness, VC tablet utilization, overall success of mission. A web application was the interface between the database and the Mission Control team. It allowed MC to observe breadcrumbs of the teams and observe how well they were fairing in making it to the desired AOIs.

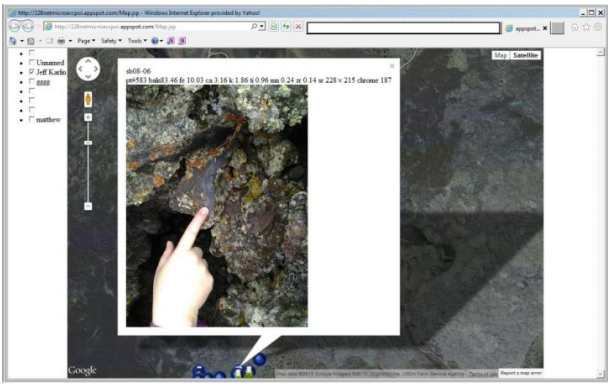


Fig. 5. A science data point entered into the VC database showing the annotation, data collected by a hand held sensor and pointing to the spot on the rock that the data was collected. View is of the web page data center.

6 Conclusions

The virtual camera for planetary exploration will assist in future exploration involving humans and robots. It is designed to provide improved exploration situational awareness and collaboration to define areas of interest and safety concern on remote planetary bodies. It acts as remote agents for mission planners and scientists capturing their knowledge and expertise as astronauts explore the solar system.

Human-centered design techniques are being applied to the development of this tool involving all potential user groups from the beginning of the design process. Prototypes have been developed and several scenarios considered and evaluated in the field.

Moving forward, the VC interaction and incrementally updated database technology can be applied to other domains such as law enforcement, disaster first responders, aviation cockpit weather displays and control room visualizations.

References

1. Boy, G.A.: Orchestrating Human Centered Design. Springer, London (2013)
2. Platt, D., Boy, G.A.: The Development of a Virtual Camera System for Astronaut-Rover Planetary Exploration. In: Proceedings of the 2012 IEA World Congress on Ergonomics, Recife, Brazil (2012), doi:10.3233/WOR-2012-0032-4532
3. Endsley, M.: Towards a Theory of Situation Awareness in Dynamic Systems. Human Factors 37(1), 32–64 (1995)
4. Millot, P., Hoc, J.M.: Human–machine cooperation: Metaphor or possible reality? In: Proceedings of ECCS 1997, Manchester, April 9-11 (1997)
5. Pacaux-Lemoine, M.P., Debernard, S.: Common work space for human–machine cooperation in air traffic control. Control Engineering Practice 10, 571–576 (2002)
6. Stanton, N., Salmon, P., Walker, G., Jenkins, D.: Genotype and phenotype schemata and their role in distributed situation awareness in collaborative systems. Theoretical Issues in Ergonomics Science 10(1), 43–68 (2009)
7. Grudin, J., Poltrock, S.: Computer Supported Cooperative Work. In: Soegaard, M., Dam, R.F. (eds.) The Encyclopedia of Human-Computer Interaction, 2nd edn. The Interaction Design Foundation, Aarhus (2013)
8. Gorman, J.C., Cooke, N., Winner, J.L.: Measuring team situation awareness in decentralised command and control environments. Ergonomics 49, 1312–1326 (2006)
9. Hutchins, E.: Cognition in the wild. MIT Press, Cambridge (1995)

Study on a Model of Flight Fatigue Dynamic Risk Index

Ruishan Sun, Wenshan Song, Jingqiang Li, and Wanli Tian

Research Institute of Civil Aviation Safety,
National Key Laboratory of Air Traffic Operational Safety Technology,
Civil Aviation University of China,
Tianjin, China
{sunrsh,happytianwanli}@hotmail.com, wenshan_song@163.com,
jqli@cauc.edu.cn

Abstract. Fatigue has been a threat to flight safety. Based on former researches about pilots' fatigue, fatigue risk and prevention measures of fatigue, a flight fatigue dynamic risk index is developed to evaluate the fatigue risk during flight operations. Flight fatigue dynamic risk index is defined as the ratio of required and human alertness during flight operations. The required alertness is the function associated with weather conditions, airport conditions, phase of flight, etc., while pilot alertness is calculated by an alertness prediction model which is based on the concept of alertness energy and constructed by human circadian oscillator, alertness energy consumption and restoration. The flight fatigue dynamic risk index prediction model should improve fatigue risk management system(FRMS).

Keywords: flight fatigue risk, safety, alertness, prediction, circadian oscillator.

1 Introduction

With rapid economic growth and social progress, the civil aviation industry of China has an enormous development. The main transportation indicators maintain steady and rapid growth in recent years, such as all civil airports completed movements 6603.2 thousand sorties in 2012, compared with 2011 there was an increase of an10.4%; all civil airports passenger capacity were 680 million in 2012, compared with 2011 there was an increase of 9.5% [1].

China's civil aviation transportation volume continues to grow rapidly, and the uneven distribution of traffic flow is mainly concentrated in a small number of political, economic and tourist center city's airports, especially in the more developed eastern of China. Statistical data showed that the passenger throughput capacity of eastern region is 389 million, accounting for 57.2% of total, only the passenger throughput of Beijing, Shanghai and Guangzhou airports account for 30.7% of all airport passenger throughput capacity [1]. The increasing civil aviation transportation results in increased pilot workload, flight fatigue phenomenon is becoming more and more common, fatigue incident occurred frequently.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 375–386, 2014.

In fact, fatigue has become a serious threat to flight safety on a global scale. On February 13, 2008, a Bombardier CL-600 of GO! Airlines performed flight from Hawaii Honolulu to Hilo. The aircraft flew over the destination and continued over 18 minutes without contacting with controllers. The investigation revealed that 'two pilots were unconsciously asleep during the cruise phase'. Until they woke up and contacted with ATC then led back safety to Hilo Airport [2]. Colgan Air flight 3407 crashed en route to Buffalo in February 12, 2009, the aircraft were destroyed, 50 people on board were killed [3]. This accident reveals a common fatigue problem among the American commuter flight crew. Because of this accident and public pressure with it, FAA forced to permit American Rulemaking Committee to develop a proposal based on human fatigue science research.

The ICAO definition of crewmember fatigue was given as: A physiological status of reduced mental or physical performance capability resulting from sleep loss or extended wakefulness, circadian phase, or workload (mental and/or physical activity) that can impair a crew member's alertness and ability to safely operate an aircraft or perform safety related duties [4] Researchers in different fields study the flight crew fatigue problem with different views. From the measured point of view, there are two kinds of method in personnel fatigue real-time measurements——subjective method and objective method. The objective method often measure some biochemical indicators to determine the people fatigue situation, commonly used biochemical indicators involving multiple systems of heart, lung, brain and so on, such as heart rate, heart rate variability, electroencephalogram, pupil diameter, etc. Although this method can obtain the accurate fatigue information of subjects, it is usually used for laboratory research now, not suitable for application in actual operations. The subjective method mainly relies on a variety of self-rating scale, determine subjects' fatigue situation through their subjective fatigue feeling. The commonly used self-rating scale include: Karolinska Sleepiness Scale, Epworth Sleepiness Scale, Samn-Perelli Scale, etc. However this method is more subjective, susceptible to subjects' emotion. From forecasting perspective, many countries commonly used biomathematics model to predict the personnel fatigue situation, but this kind of model is considered to be imperfect and temporary scientific instruments, those models developed by each country also have some limitations and shortcomings.

The flight fatigue research group in the Research Institute of Civil Aviation Safety of Civil Aviation University of China considered that the most direct impacts on the flight safety of fatigue are decreasing the alertness of pilots. The study constructed biomathematics to predict the pilot's alertness. The group used the alertness to describe the pilot's fatigue status from operational angel, and constructed the prediction model of flight fatigue dynamic risk index.

2 Biomathematics Model on Fatigue

In the field of fatigue prediction, there has been a lot of research achievements, including the human physiological parameters associated with the body fatigue for the data source, through the mathematical formula of quantitative calculation mathematical model of biological fatigue risk. Mathematical Model for Fatigue life

from the initial Two Process Model (Two Process Model of Sleep Regulation, TPMSR) [5-6] began has developed more than 30 years, spawned many new models of different characteristics, such as Fatigue Audit InterDyne, FAID [7], System for Aircrew Fatigue Evaluation, SAFE [8], Circadian Alertness Simulation, CAS) [9], etc.

In 1987, Torbjorn Akertedt expanded on the basis of TPMSR and put forward Alertness Three Process Model (ATPM), also known as the Sleep/Wake Predictor, SWP. The Model believed that the factors that affect Sleep and awakening time included not only the homeostatic Process S and the circadian Process C but also the wakeup Process W meaning Sleep inertia [10], Now many of the mathematical models for Fatigue life are established on the basis of the three processes, such as SAFE, CAS, Fatigue, Sleep, Activity and Task efficiency (Sleep, Activity, Fatigue and Task Effectiveness, SAFTE) model [11]. Each model has its own advantages and disadvantages, carries on the contrast analysis as follows.

(1) SAFE is a fatigue risk management tool, which is developed by QinetiQ with the support of the civil aviation authority and used to help airlines crew to assess fatigue risk due to shifts. Model was established based on the laboratory data at first, then according to the short, long distance and cross intercontinental long-haul flights research data the model was calibrated.

Advantages: the model according to the actual operating environment of airline has been corrected, including the landing airport code and time zone, flight segment number, the aircrew consisting and resting place (bed, sleeping on the plane, cockpit seats) and other factors. And as a result it made the model more suitable for the aviation industry, taking into account the special nature of the aircrew career and route length of distinction and in the long range and intercontinental voyage it added time zone changes and prolonged operation factors.

Disadvantages: individual differences were not considered in.

(2) CAS was originally based on steady status process and circadian rhythms to portray individual alertness curve, after three relatively mature process development it will also take sleep inertia into account. With the development of the research, CAS is also in constant progress, now it has been updated to CAS - 5.

Advantages: the model has been applied to some airlines' actual operation, including the pilot ID, airport code, duty start/end time (briefing before the flight, flying, reports after the flight, setting segment, etc.) and other characteristics of civil aviation. Sleep also is given more consideration and sleep pattern can be divided into early in the morning or evening, short/long sleepers, sleepiness type, etc.

Disadvantages: its validity only verified in the ground transportation industry.

(3) SAFTE is a Fatigue prediction model developed by SAIC (Science Applications International Corporation) and NTI (Network Technologies, Inc.) with the support of the U.S Air Force in 2001. The model was built up on the basis of laboratory data. This model took sleep reservoir as the center and sleep is divided into own sleep regulation and flight performance adjustment two parts.

Advantages: SAFTE specified the steady-status process and took into account the impact of sleep quality. If poor sleeping environment resulted in sleep disruption, the actual sleep starting time should be delayed five backwards minutes after falling asleep again.

Disadvantages: the model for the calculation of the period of sleep consume is too simple and all the consumption of nature of work is exactly the same.

Most of the fatigue prediction models are based on sleep deprivation experiments, namely the prediction of fatigue is established by two parts' correlation, and the two parts are individual subjective fatigue due to lack of sleep/under the limit and awareness. These models believed that awareness changed during pilots' working process mainly because of sleep time, sleep quality and its circadian rhythm characteristics between the roles of the model. However, from the angle of the operation, the model still has many places for improvement.

First, the design and development of the model should be combined with industry characteristics: the job of aircrew is different from other industries and many models did not specifically target in civil aviation in the beginning of establishing. Although the individual model according to the airlines' relate research is improved, including the airport code, and the factors such as travel, but still lack of validation of real air operating environment.

Second, the models should predict the flight fatigue risk in an accuracy way which is one of the significant means to ensure the safe operation in flight. The model is mainly based on laboratory data which is set up for most people in common situations, but there are individual differences between human beings such as age, job experience and so on, there is a certain gap between the individual actual risk of fatigue and models' predicting results . Actual operating environment is complicated and cannot take all the factors into account, such as airport complexity, weather conditions and other important factors affecting safety, and previous models do not take these into account.

3 The Basic Idea of Constructing Flight Fatigue Dynamic Risk Index Prediction Model

The study aimed to study on fatigue risk during flight operations. Fatigue makes human's impairment in attention and alertness, leading to human errors, which is the most direct impact on flight safety.

The definition of flight fatigue dynamic risk index is given as: the ratio of required and one's alertness during flight operations. The required alertness is the function associated with weather conditions, airport conditions, phase of flight, etc.

The focus of study is constructing alertness prediction model, which presents the concept of alertness energy, proposing restoration and consumption of alertness energy and improvement of alertness on the basis of SAFTE model for reference.

Alertness is given by the following equation:

$$E = 100*(\frac{Rt}{Rc})+C+I+B \tag{1}$$

Where Rt= current alertness energy level, Rc= one's alertness energy threshold (maximum), C= alertness rhythm, I= sleep inertia and B= enhancement effect of the alertness.

Alertness rhythm is represented by the following equation:

$$C = \mathrm{Ap} * c \tag{2}$$

Where Ap= alertness rhythm amplitude, and c=circadian oscillator.

Alertness rhythm amplitude is represented by the following equation associated with Rt/Rc and age:

$$\mathrm{Ap} = f\left(\frac{Rt}{Rc}, \mathrm{z}\right) \tag{3}$$

Where z= age.

Circadian oscillator is represented by the following equation:

$$\mathrm{c} = \cos(2\pi(T-p)/24) + \gamma\cos(4\pi(T-p-p')/24) \tag{4}$$

Where T= time of day, p=24 hr phase in hours, p'=12 hr relative phase in hours and γ=relative amplitude of 12 hr cycle.

Current alertness energy level is represented by the following equation:

$$\mathrm{Rt} = \int SI * \mathrm{Sq}dt - \int Kdt \tag{5}$$

Where SI= sleep intensity, Sq=sleep quality and K= alertness energy used rate.

Sleep intensity is represented by the following equation associated with current alertness energy deficit and circadian oscillator:

$$SI = f(\mathrm{Rc} - Rt, c) \tag{6}$$

Where Rc-Rt= current alertness energy deficit.

Sleep quality is represented by the following equation associated with self-evaluation of sleep quality, sleep conditions and snoring:

$$\mathrm{Sq} = f(Sf, Sc, Sd) \tag{7}$$

Where Sf=self-evaluation of sleep quality, Sc=sleep conditions and Sd=snoring.

Alertness energy used rate is represented by the following equation associated with type of duty, numbers of sectors, experience, airport conditions, etc.:

$$K = f(D\mathrm{t}, f\mathrm{n}, f\mathrm{e}, fa) \tag{8}$$

Where Dt=type of duty, fn=numbers of sectors, fe=experience and fa=airport conditions.

3.1 Alertness Supplement

Sleep is an effective measure to mitigate fatigue and increase alertness energy, but poor quality of sleep will make it difficult to recover from fatigue. Sleep quality is

affected by sleep environment, sleep hygiene, individuality, etc. In this section, we will analyze influencing factors of sleep quality in detail.

Sleep quality. Sleep quality measurement and evaluation methods are roughly divided into two categories including subjective and objective measurement. Subjective measurement includes filling sleep logs, questionnaires, surveys, etc. And objective measurement includes polysomnogram (PSG) and so on [12]. However, objective measurement which is not suitable as inputs to the model demands high standards on device and experimental conditions, so the model uses the subjective measurement to evaluate pilot's sleep quality as a corrected parameter.

Sleep condition. The various sleep conditions including sleep environment and sleep patterns are closely associated with sleep. The issues affecting the sleep environment including temperature, shading, sound insulation. The most suitable sleep environment is 20-23℃ of the room temperature. When the noise exceed 35 db or strong light would disturb sleep obviously [13]. However, objective measurements of the temperature, shading and sound insulation of the sleep environment are too complex. To simplify the system input, subjective self-evaluation is used to evaluate the sleep environment.

As the civil aviation industry characteristics, pilots often fly across the region. The operators often arranged pilots to sleep overnight in other places to meet the rest requirement of regulators. Given the form of sleep patterns would affect the quality of sleep, the patterns are classified as habitual residence beds, temporary shelter beds and floor seats.

Individuality. There are big difference among the sleep quality of different individuals as there are different individual snoring circumstance and sleep patterns. In order to evaluate personal alertness in the model and ensure the maximum accuracy and minimum uncertainty, the effects of personal characteristics such as snoring and sleep patterns are considered.

Snoring is a common sleep breathing disorders, which is due to the vibration of the soft palate oropharynx when the air goes through the oropharynx. Snoring means that there are some narrowing and blocking in airway, which would increase acting when breathing and cause awakening during sleep, resulting sleep fragment, and affecting the quality of sleep [12].

Judging from the sleep patterns, some people are morningness and others are eveningness. Some people are accustomed to have napping while others are not. Because of the different scheduling, pilots get variable daily sleep schedules and different sleep patterns would affect the quality of their sleep.

Sleep quality evaluation indicator system are initially established through the analysis of existing theories, shown in Figure 1, including the self-assessment of sleep quality, sleep conditions and personal characteristics, sleep quality in the model is expressed by the functions associated with these indicators.

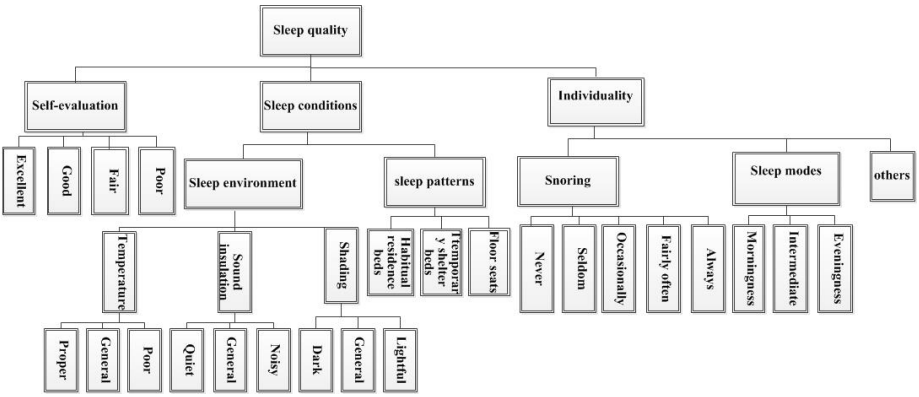


Fig. 1. Sleep quality evaluation indicator system

3.2 Consumption of Alertness

Continuous awakening would reduce human alertness. The alertness would be consumed when awake and workload would also reduce the alertness. Both of these are considered in the system. Workload refers to the unit of time the body to withstand the workload, which including the effects of environmental conditions, social and psychological factors on the human body [14]. The higher workload at work, the greater the energy consumption rate, the more obvious fatigue that people feel and its time sooner.

$$K = Ka + Kw \tag{9}$$

Where Ka= consuming rate in awake status and Kw= workload consuming used rate associated with numbers of sectors, experience, airport conditions, etc.

Types of working duty. In term of pilots, types of the working duty include flight duty, simulator instructing, management, transition and so on, which leads to different workload consuming rate. In 2002, the UK Civil Aviation Authority completed a study on aircrew alertness during short-haul operations. The study was conducted by analyzing the collected diaries to identify factors that contributed to a buildup in fatigue. The diary used a modified 7-point Samn-Perelli scale to evaluate pilot's subjective fatigue level. The study showed that fatigue levels of pilots on flight duty represent an increase of 0.16 per hour on the 7-point fatigue scale than those of pilots on rest. When compared to the output of the model, was equivalent to a 2.656 decreasing alertness per hour. The alertness use is represented by a linear function [15], so the consuming rate of workload of flight duty Kf is 0.04.

Numbers of route segment. A study has shown that short-haul operations cause frequency times of taking off and landing, increasing the workload. This means that pilots with multiple segments are especially prone to fatigue. Pilots once reported that it is more tried during short-haul operations with 4-5 sectors [17]. The increase in

fatigue from one to four sectors is much closed to linear. An additional sector is equivalent to an additional 45 minutes' flight duty [16]. So the workload used rate in the nth sector is represented by the following equation:

$$Kf(n)=\begin{cases} Kf & n=1 \\ \dfrac{(Ka+Kf)*(TLn+45)}{TLn}-Ka & n>1 \end{cases} \tag{10}$$

Where n= numbers of sectors and TLn= flight time in the nth sector.

Working experience. An experienced pilot is more familiar with operating procedures. If the pilot operates the airplane with a proper way and a reasonable time planning, it will largely reduce the workload and delay his fatigue status. The more flight hours the pilot has, the less workload the duty is.

After all these comprehensive consideration about the effect of duty type, flight phase, experience and so on, the workload consuming rate is represented by the following equation:

$$Kw=\begin{cases} Ks & \text{simulator instructing} \\ Km & \text{management} \\ f_a*F*Kf(n) & \text{takeoff} \\ F*Kf(n) & \text{other flight phases} \\ f_a*F*Kf(n) & \text{landing} \\ f_x*Kp & \text{briefing} \\ f_x*Kt & \text{transition} \\ Kb & \text{debriefing} \\ Kc & \text{commuting} \end{cases} \tag{11}$$

$$F=w_b*f_b+w_e*f_e \tag{12}$$

Where Ks=simulator instructing used rate, Km= management used rate, F= influence coefficient of flight duty, fa=airport factor (fa(complexity, general)=(1.1, 1)), fx=flight nature factor(fx(delay, on-time, diversion)=(1.05,1,1)), Kp= briefing used rate, Kt= transition used rate, Kb= debriefing used rate, Kc= commuting used rate, fb=position factor(fb(captain, copilot)=(1, 0.9)), fe=experience factor(fe(flight hours: 0-800,800-1500,1500-3000,>3000)=(1,0.95,0.90,0.85)), wb= the weight of position factor and we= the weight of experience factor.

3.3 Alertness Rhythm

A circadian rhythm is a daily alteration in a person's behavior and physiology controlled by an internal biological clock located in the brain. With the rhythm changes of a person's physiological and psychological status, alertness also appears

rhythm changes. Moreover, Alertness rhythm amplitude is associated with current alertness energy level and age.

A study showed that a person's tissues and organs functions were reduced with the increase of the age, leading to flat amplitude of rhythm. Ge Shengqiu used critical flicker frequency (CFF) test and cross out test to study effect of flight fatigue in aircrew of different age groups. The result showed that The performance in aircrew ≤40 was significantly higher than that of 41～50 , > 51 age groups , but the difference between 41～50 age group and> 51 age group was not significant[17]. Therefore, amplitude of rhythm in aircrew ≤40 and > 40 age group is different, and older aircrew's amplitude is lower.

3.4 Alertness Improvement

Fatigue problem will be a serious threat to flight safety, so staff alertness should be effectively improved through some mitigation measures. For example, the pilot can drink some drinks containing caffeine to promote alertness; Pilots can also rest between airlines to take a nap, or during long flight of more aircrew nap can alleviate fatigue. But for a nap before flight or in the flight, at the same time, sleep inertia caused after the nap needs special attention.

Caffeine. Caffeine is a central nervous system stimulant. It can temporarily fade away sleepiness and enhance alertness when humans drink caffeine at low doses (50mg-200mg). The average cup of coffee contains 50-150mg caffeine. Tea, chocolate, cocoa and many other drinks are the common source of caffeine. A study showed that peak blood concentration is reached in 30 minutes to one hour, and its half-time is 3-5 hours [12]. Alertness enhancement effect of caffeine is represented by the following piecewise equation:

$$B = \begin{cases} \alpha * \sin \dfrac{\pi * t_c}{2 * t_g} & 0 \le t_c < t_g \\ \alpha e^{-(t_c - t_g)/\beta} & t_c \ge t_g \end{cases} \tag{13}$$

Where α=maximum alertness enhancement effect of caffeine, β= caffeine's half-time, tg= time when reach the peak alertness enhancement effect of caffeine and tc= minutes since intake of caffeine.

Nap in cockpit. A research showed that reasonable arrangements of naps can reduce the sense of subjective sleepiness and fatigue of pilot. In the long-haul operations or a continuous flight duty in special circumstances, , reasonable arrangements of naps can relieve pilots' fatigue. But it should be noted that naps can lead to sleep inertia. After awakening from sleep, before desired level of alertness there will appear a delay [18]. To reduce the risk of sleep inertia caused by nap, it is recommended that nap time limits should within 40 minutes. In the model, pilots cannot nap in the takeoff and landing. Forms of napping in the cockpit include bunk and seat.

3.5 The Alertness Required for Each Flight Phase

During the flight operations, different phases of flight require different alertness. Takeoff and landing which called "dangerous 11 minutes" are recognized as phases with biggest workload. Boeing statistical data showed that most of aviation accidents occurred in these 11 minutes (Figure 2) [19]. Therefore, pilot requires higher alertness during the takeoff and landing. If lower alertness during takeoff and landing phases will result in a higher safety risk.

According to how busy airports are and geographical conditions of the airport, the airports are divided into complex airports and general ones. Complex airports will require pilots' higher alertness during takeoff and landing .In addition, the airline weather, early starts or night flight are major factors affecting flight workload. The required alertness is represented by the following equation:

$$Er=\begin{cases} J*fa*fw*fd & \text{Takeoff} \\ J*fw*fd & \text{Climb,cruise,descent} \\ J*fa*fw*fd & \text{Approach,landing} \end{cases} \tag{14}$$

Where J= the required alertness for each phase of flight(J（takeoff, climb, cruise, descent, approach and landing）=（75,70,70,70,80）), fa=airport factor (fa(complexity, general)=(1.1, 1)), fw= complexity of the airline weather (fw(complexity, general)=(1.1, 1)) and fd=nature of duty (fd(early start, day flight, night flight)=(1.1, 1, 1.2)).

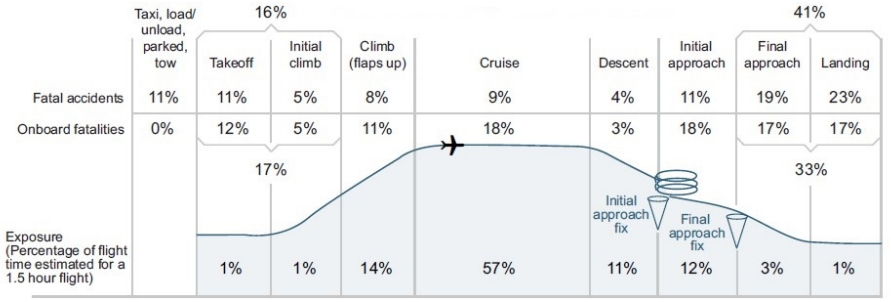


Fig. 2. Percentage of fatal accidents

3.6 The Fight Fatigue Dynamic Risk Index

The fight fatigue dynamic risk index is the ratio of the predicted alertness and the alertness required for each flight phase. The safety risk of different fight phases are assessed through the required alertness and the predicted one. The larger index indicates a higher risk, and the smaller index indicates lower risk. If the dynamic risk index is less than 0.9, it indicates that the pilot's fatigue risk is acceptable; If is between 0.9-1, it indicates that the fatigue risk is in the tolerable range, some measures like naps or drinking coffee should be taken for mitigation; when is greater

than 1, it indicates that the fatigue risk is unacceptable, some measures like the replacement flight crew should be taken for mitigation. The fight fatigue dynamic risk index is represented by the following equation:

$$I_E = \frac{E\mathrm{r}}{E} \tag{14}$$

Where Er=the required alertness and E= the predicted alertness.

In summary, based on index in different flight phases, we can acquaint the fatigue risk of pilots in operations. When the risk is striking, we can increase resting time, shifting, napping in the cockpit, drinking coffee and so on. In this way, the safety risk can remain within an acceptable range.

4 Prospects

With the rapid development of air transportation, staff shortage and flight delay phenomenon are getting worse, leading to staff's fatigue phenomenon, especially for pilot, have been more and more common in front-line of civil aviation. So the reduction, control and management of flight fatigue have become an important work for improving the flight safety. The model of flight fatigue dynamic risk index is comprehensively integrated of many factors of fatigue and expressed by the simple concept of alertness energy and flight fatigue dynamics risk index, which is advantageous in spreading and application. The model should be a support for implementing fatigue risk management system(FRMS).

Acknowledgments. We appreciate the support of this work from the National Natural Science Foundation of China (No. 61304207, No.U1333112), the National Basic Research Program of China (No.2010CB734105), and the Science and Technology Funding of CAAC (No. MHRD2011238).

References

1. The development of statistical bulletin in civil aviation industry in 2012 (2012), http://www.caac.gov.cn/I1/K3/201305/P020130520548774552650.pdf
2. Werfelman, L.: The Science of Fatigue. J. AeroSafety World, 40–43 (2008)
3. Accident Description, http://aviation-safety.net/database/record.php?id=20090212-0
4. International Civil Aviation Organization: Fatigue Risk Management Systems Manual for Regulators (2012)
5. Achermann, P.: The two-process model of sleep regulation revisited. J. Aviation, Space, and Environmental Medicine 75(3), A37–A43 (2004)
6. Achermann, P., Borbely, A.A.: Simulation of daytime vigilance by the additive interaction of a homeostatic and a circadian process. J. Biological Cybernetics 71, 115–121 (1994)

7. Roach, G.D., Fletcher, A., Dawson, D.: A Model to Predict Work-Related Fatigue Based on Hours of Work. J. Aviation, Space, and Environmental Medicine 75(3)(suppl.), A61–A69 (2004)
8. Fatigue Risk Management Systems Limited. The SAFE Model. In: FRMS Forum, Montreal (September-December 2011)
9. Chick, S., Sanchez, P.J., Ferrin, D., et al.: Human Fatigue Risk Simulations in 24/7 Operations. In: Proceedings of the 2003 Winter Simulation Conference, pp. 1838–1842 (2003)
10. Torbjorn Akerstedt, S.F., Portin, C.: Predictions from the Three-Process Model of Alertness. Aviation, Space, and Environmental Medicine 75(3)(suppl.), A75–A83 (2004)
11. Hursh, S.R.: System and Method for Evaluating Task Effectiveness Based on Sleep Pattern: US,6579233 B2 (June 17, 2003)
12. Kryger, M.H., Roth, T., Dement, W.C.: Principles and Practice of Sleep Medicine. Saunders/Elsevier, Netherlands, Amsterdam (2011)
13. Zan, H., Chen, Y.S.: The Reserch on Flight Fatigue, pp. 139–141. National Defense Industry Press, Beijing (2011) (in Chinese)
14. Zhu, Z.X., Ge , L.Z., Zhang, Z.J., et al.: Engineering Psychology, pp. 269–332. People's Education Press, Beijing (2000) (in Chinese)
15. Hursh, S.R., Redmond, D.P., Johnson, M.L., et al.: Fatigue Models for Applied Research in Warfighting. J. Aviation, Space, and Environmental Medicine (3)(suppl.), A1–A10 (2004)
16. Ge, S.Q., Wu, G.C., Xu, X.H., et al.: Effect of flight fatigue on critical flicker frequency in airline aircrew of different age groups. J. Chin. J. Aerospace Med., 180–183 (2005) (in Chinese)
17. Ward, P., Spencer, M.B., Robertson, K.A.: Aircrew alertness during short-haul operations, including the impact of early starts. Technical report, the UK Civil Aviation Authority (2001)
18. Dinges, D., Orne, M., Orne, E.: Sleep depth and other factors associated with performance upon abrupt awakening. J. Sleep Res. 14, 92 (1985)
19. Commercial Jet Airplane Accidents, `http://www.boeing.com/news/techissues/pdf/statsum.pdf`

Safety Culture Evaluation in China Airlines: A Preliminary Study

Chiou-Yueh (Judy) Tsay[1], Chien-Chih Kuo[2], Chin-Jung Chao[3], Colin G. Drury[4], and Yu-Lin Hsiao[5]

[1] China Airlines, Taoyuan County 33758, Taiwan
mdprincess@china-airlines.com
[2] Department of Psychology, National Chengchi University, Taipei City 11605, Taiwan
cckuo@nccu.edu.tw
[3] Department of Industrial and Systems Engineering,
Chung Yuan Christian University, Chung Li 32023, Taiwan
{davidchao,yhsiao}@cycu.edu.tw
[4] Department of Industrial and Systems Engineering,
State University of New York (SUNY) at Buffalo, Buffalo, NY 14260, USA
colindrury@hotmail.com
[5] Department of Industrial and Systems Engineering,
Chung Yuan Christian University, Chung Li 32023, Taiwan
yhsiao@cycu.edu.tw

Abstract. Recently, the global aviation industry has started to promote Safety Management System (SMS), and thus consider the enhancement of safety culture as an essential. To meet the growing demand of safety culture assessment, we cooperated with China Airlines (CAL) in Taiwan to develop a safety culture questionnaire using both qualitative and quantitative methods to meet the needs of their operation environment and to take the theoretical researches of safety culture into account. During the development process, we continuously integrated the opinions from the industry experts, and effectively established a safety culture assessment tool that not only conforms to the actual operation of the airline but also to the requirement of reliability and validity. Although the statistical results show both the strengths and weaknesses of CAL's safety culture, the hidden reasons of the shortcomings or unstated ideas of workers are still remained undiscovered in this phase. In the following, we would use the focus group method to help examining the potential causes of those low score facets of safety culture, and to develop follow-up recommendations to improve CAL's safety performance accordingly.

Keywords: Aviation, SMS, Safety Culture, Reliability, Validity.

1 Introduction

Despite the aviation technology have obtained significant progress since the early twenty century and flight accidents for design and mechanical factors have substantial reduced correspondingly, it is unfortunately that people who use and maintenance aircraft don't evolve with the advancement of technology. As modern people, we still have no difference to our ancestors on cognition, decision-making, action, and so on.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 387–397, 2014.

According to the safety report of International Air Transport Organization (IATA), about 70% of the flight accidents are caused by human factors in these two decades. Without any doubt, human factors do play an unreplaceable role in contemporary flight safety management. The field not only include the behavior and performance of first-line workers (e.g., pilots, maintenance technician, air traffic controllers, dispatcher, and so on), but also covers the impact of the environment, management, and the organization function. Among these, one of the most popular issues is the influence of organizational safety culture in recent years.

Safety culture was first appeared in the accident report of Chernobyl nuclear power plant by the International Atomic Energy Agency (IAEA, 1992). The report concluded that no sound safety culture led to the operators' error and thus was one of the causes of the Chernobyl accident. Since then, the safety society starts aware that the overall safety culture of an organization has a significant effect on the safety performance of the staff and the organization itself. If employees of an organization share a consensus such as "safety first", the chance of human error can be reduced, and thus safety can be improved consequently.

Over these years, researchers and safety-related industries have conducted lots of safety culture studies. Although we still argue about the definition and the explicit composition of safety culture, most have agreed that compared to the safety policy and other "official" safety regulations, safety culture is the hidden or "unofficial" rules of an organization. It is "the way things get done around here", and is the shared values, principles, attitudes, and traditions in an organization. Many studies have proven that safety culture is usually the latent cause of human error, incident and accident, and have long-term impact on flight safety. Having a proactive safety culture will bring positive benefits to the long-term safety performance of an organization.

In the aviation industry, ICAO has advocated Safety Management System (SMS) step by step for years asking all airlines to establish their own SMS mandatory. In their Safety Management Manual, ICAO has emphasized the impact, progress and recommendations of safety culture. Under this trend, the airlines in Taiwan have gradually paid more attentions to their own safety culture, and have begun to seeking valid ways to enhance it in the long-term.

In this study, in order to effectively assess the status of safety culture and to take account of national characteristics and conditions, we cooperate with China Airlines (CAL) to get their full support. We used both qualitative and quantitative methods such as pilot test, reliability, validity analysis and expert panel discussion during the developing progress. The attempt is to establish a suitable safety culture questionnaire to examine CAL's safety culture, and to utilize the results to help developing appropriate recommendations for their culture improvement.

2 Literature Review

Safety culture might be slightly different because of in which nation, region, industry and corporate. Although the researchers of the safety management field might argue the definition of safety culture and deliver different description about it (Geller 1994, Clake 1999, Cooper 2000), the same purpose is hoping make a more explicit interpretation of safety culture.

To develop a practical safety culture assessment tool to meet CAL's need, we have reviewed related safety culture studies to clarify the dimensions of safety culture, and have combined our thoughts and the airlines' professional opinions to propose our own safety culture framework.

In this study, we simply defined safety culture as "the common characteristics of the organizations and individuals that is composed by the interactions between various facets of personal, work and organization." It will have a comprehensive impact on the cognition, perception and safety behavior of all members at work environment. Since safety culture involved many related components in the working conditions, in the management practices, and in the organizational functions, it is important to further understand what issues should be taken into account in safety culture model.

Many organizations have investigated historical events from safety culture perspective to understand the reasons for the occurrence, and to improve its deficiencies as well. Government agencies, industries and researchers have continuously committed to the development of a comprehensive and applicable safety culture model, and have expected to use it to enhance the effectiveness of safety culture (Cole, Stevens-Adams, & Wenner, 2013).

In total safety culture model, Geller (1994) has proposed that safety culture consists of individual, environmental and behavioral aspects. The individual dimension refers to the psychological and mind state such as knowledge, skills, abilities, intelligence, motivation and personality. The environment dimension is regarded as hardware, tools and SOPs. And the behavioral aspect is referred to as the conduct of safe performance, and comprises regulation compliance, hazard identification, communication and so on.

Reciprocal model of safety culture is based on the social cognitive theory, and Cooper (2000) extended this concept into safety culture. The model is composed of the interactions of the individual psychological, behavioral and organizational level. Cooper argued that safety culture can be measured based on the three key dimensions. First, we could use safety climate survey to assess the safety attitudes and perceptions of individuals, conduct behavioral observation to evaluate the behavioral level, and then have an objective way to review the organizational safety performance.

On the other hand, HSE (2005) has conducted a literature review and developed a safety culture assessment framework for railroad transportation industry. The framework includes leadership, two-way communication, employee involvement, learning culture, and punitive attitude as the five core factors of safety culture. The various facets of safety culture, and the practical methods and guidelines to promote safety culture have been described in details in the HSE report. It provides great values during the development of our safety culture model.

Based on the progress of airlines' SMS in Taiwan, Liou et al. (2008) claimed that using current accident rates to predict future safety performance is no longer the best way. A more forward-looking approach is applying the human factors concept to assess the safety factors that are related to the staff. Their research is based on the triangle architecture of safety management system (McDonald, Corrigan, Daly, & Cromie, 2000) that includes organizational strategy and policy, personal factors (ability, training and communication), and execution (working drills, equipment and

document). Liou et al. consider a sound safety management system should be able to effectively promote three above influencing factors, and thereby reducing the probability of flight accidents.

Cox, Cheyne, and Alexander (1997) developed a safety culture model that includes individuals, organizations and behavior dimensions. The unique advantage of this model is the emphasis on the impact of the organizational management. They believe that management attitude and commitment to safety will affect the subordinates' safe behavior. Although Cox and Liou both address the influences of organization level, Cox points out in particular that behavior and promises of managers are more important than promotion of good safe behavior.

After the integration of these proposed safety culture framework, we have developed a HF-centered safety culture model. The overall framework covers four levels: organizational system, executive, immediate supervisor and staff.

The organizational system consists of safety policy, safety management system and organizational functionality and resources. We divided the management into top and immediate managers based on their different roles and management functions. The high-level executives would only focus on their commitment to safety, and for the immediate supervisor, we would like to evaluate their safety-related activities and attitudes. The last but not least, the staff level includes the safety attitudes and communication of themselves and colleagues, and safety-related context at work place.

The simple framework is shown in Figure 1. We argue that in an organization with poor safety culture, the influence flow between the four levels would be top-down only, but with a proper safety culture, both the communication and influence flow should be two ways. Conducting a safety culture survey would be a good progress to bring a consensus to both managers and employees and to enhance safety culture consequently.

Fig. 1. Safety culture framework

3 Questionnaire and Method

Our safety culture model is referred to the relevant literature and based on the opinions of CAL's experts and the operational characteristics of CAL. The selected

factors which were included in the model are considered to be effective to characterize the dimensions of safety culture in previous studies.

Basically, this framework is found on the organization behavior concepts, and we try to generalize the relevant elements with safety culture from organization, management, and individual perspectives. The detailed composition of the safety culture questionnaire is shown in Table 1.

Table 1. Safety culture mode architecture table

Level	Sublevel	Element
Organizational System	Safety Policy	Formal safety policy Long-term safety & risk management plan Safety performance indicator (SPI) management Total involvement
	Safety Mgmt. Sys.	Self-audit Safety Reporting System Manual / SOPs / checklist Safety Training Recognition & disciplinary system Performance evaluation & promotion criteria
	Org. Function and Resources	Departmental responsibility Departmental interaction Safety investment & resource allocation
Executive	Safety commitment of top managers	Safety value of top executive
Immediate supervisor	Safety mgmt. of immediate supervisor	Safety attitude of immediate supervisor Control of immediate supervisor Conflict and performance management Communication of immediate supervisor Safety information exchange
Staff	Safety attitude & communication	Personal safety attitude Colleagues' safety attitude Interaction with colleagues
	Safety perception	Power distance Job satisfaction Workload & fatigue Risk perception Change & self-adjustment

The questionnaire has a total 27 elements with 78 questions. We have followed the steps below to establish a verified assessment tool.

1. Conduct literature review to establish a draft framework.

2. Develop question database and conduct first expert panel discussion to discuss and screen all questions.
3. Conduct pilot test, reliability analysis, and second expert panel discussions to examine the interpretative meaning and value of each question.
4. Based on the quantitative and qualitative results of step 3, establish the final version of the safety culture questionnaire.

The subjects of the survey are the staff and immediate supervisor from CAL's flight operation (OZ) and maintenance (EMO) departments. We used random stratified sampling with a sampling ratio 40% per unit, and the estimated sampling size is about 1,230.

Because subjects might be equivocal with some sensitive questions and decide to answer neutral options, we actually chose the Likert six scales rather than five to eliminate the possible ambiguous option and to collect the subjects' opinions explicitly. The six options are ① "strongly disagree", ② "do not agree", ③ "somewhat agree", ④ "somewhat agree", ⑤ "agree", ⑥ "strongly agree" respectively. The higher the answer, the more agreement of the question statement.

4 Analysis

There are total 1,245 valid samples, and the effective response rate is about 93%. The actual sampling percentage of the OZ and EMO departments both effectively reached the goal of sampling rate, 40%.

Most of the responses were from senior employees with seniorities are usually more than six years (OZ: 77% and EMO: 80%). The summary of the background information is shown in Table 2, and most background statistics are close to the overall population. This result effectively reflects the overall ideas from the employees about CAL's safety culture.

For reliability, the Cronbach's Alpha of the overall framework is 0.971. In Table 3, we also checked all sublevels' alpha values to check if their reliabilities still meet the acceptance level, 0.70. Except "Safety commitment of top managers" is not available for reliability analysis, the Cronbach's Alpha values of all sublevels are between 0.734~0.931.

Table 2. Summary of subjects' background

Flight Operation (OZ)			Maintenance (EMO)		
Age	25-34	24%	Age	25-34	12%
	35-44	44%		35-44	46%
	> 45	28%		> 45	40%
Seniority	< 5 yrs.	18%	Seniority	< 5 yrs.	8%
	6-10 yrs.	34%		6-10 yrs.	18%
	> 10 yrs.	42%		> 10 yrs.	70%
Fleet	A	34%	Grade	<= Grade7	20%
	B	30%		Grade 8-10	62%
	C	28%		>= Grade 11	16%
Pilot Training Background	Self-training	28%	Position	Manager	8%
	Company	48%		Foreman	56%
	Military	18%		Staff	32%
Flight Hour	< 3000 hours	16%	License	No	32%
	3-6000 hours	22%		Yes	66%
	6-9000 hours	22%	Work shift	Yes	46%
	> 9000 hours	34%		No	52%

The correlation analysis is conducted to see if the data is consistent between OZ and EMO. As Figure 2, it is clear that the answers from two departments are highly correlated with a correlation coefficient 0.87.

Table 3. Results of reliability analysis

Level	Sublevel	No. of questions	Cronbach's Alpha
Organizational Sys.	Safety Policy	7	0.824
	Safety Mgmt. Sys.	26	0.914
	Org. Function and Resources	6	0.774
Executive	Safety commitment of top managers	2	N/A
Immediate supervisor	Safety mgmt. of immediate supervisor	17	0.931
Staff	Safety attitude & communication	11	0.812
	Safety perception	9	0.734

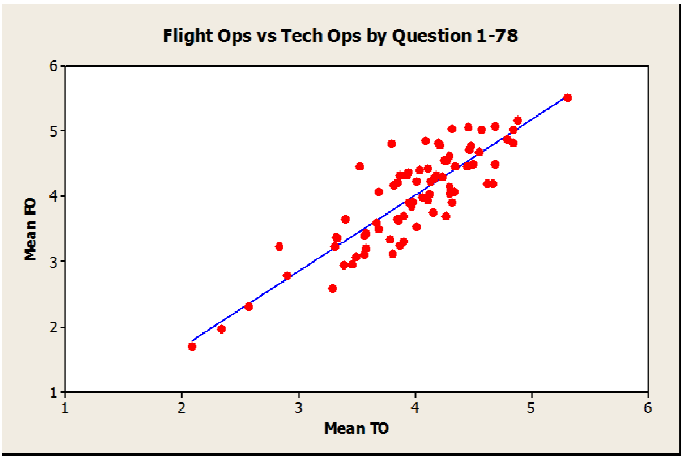


Fig. 2. Correlation result between Flight Operation and Maintenance Departments

The overall means of OZ and EMO were 3.995 and 4.09 respectively. Thus, using 4.0 (somewhat agree) as the standard value to evaluate the agreement of all sublevels and elements is a reasonable option. In Fig. 3, we found the average scores of four out of seven sublevels are close or lower than 4.0, and these sublevels were "Safety Management System", "Organization Function and Resources", "Safety commitment of top manager", and "Safety management of immediate supervisor".

We listed the mean of all elements in Fig. 4 to scan both the strengths and weakness of CAL's safety culture. The scores of nine out of 27 elements were below 4.0, and eight out of the nine low scored elements belong to the four sublevels whose means were lower than 4.0 except the "Workload & fatigue" element (see Table 4).

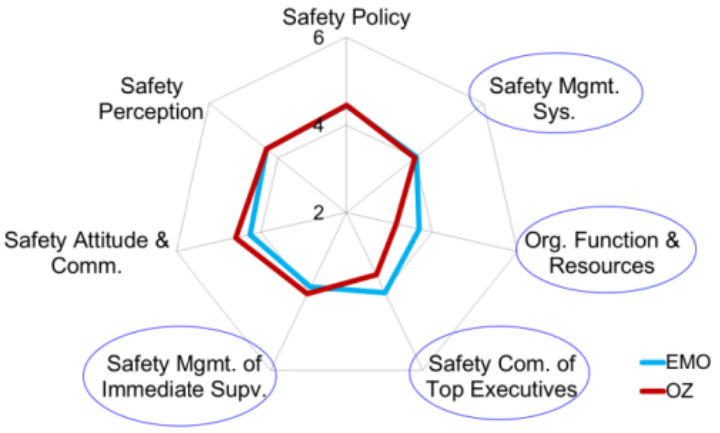


Fig. 3. Scores of the sublevels of the safety culture model

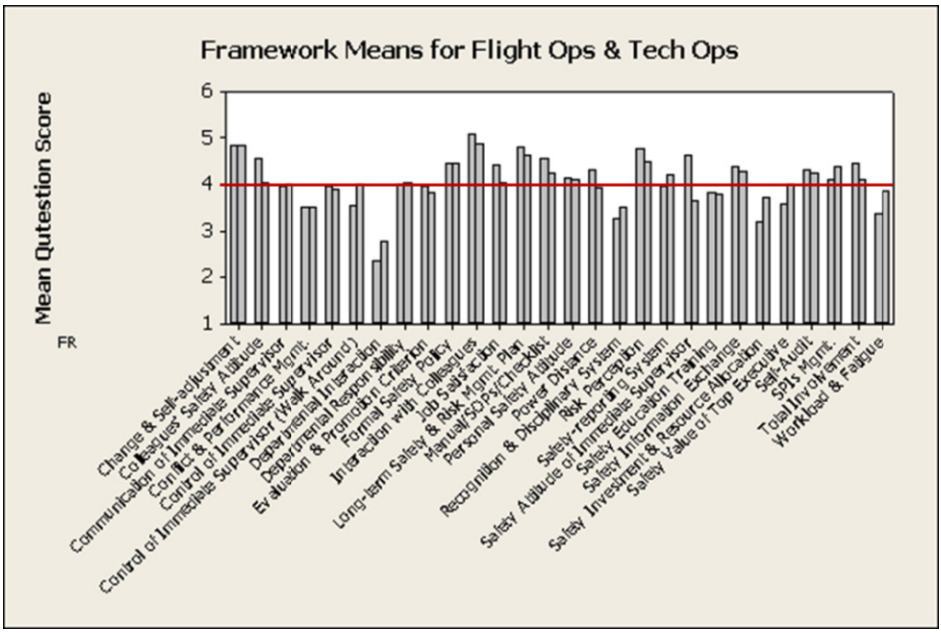


Fig. 4. Overall scores of the elements of the framework

Table 4. The low scored sublevels and elements

Sublevel	Element	EMO	OZ
Safety Management System	Recognition & disciplinary system	3.67	3.06
	Performance evaluation & promotion criteria	3.86	3.94
Organization Function and Resources	Departmental interaction	3.29	2.34
	Safety investment & resource allocation	3.75	3.17
Safety commitment of top manager	Safety value of top executive	4.04	3.57
Safety management of immediate supervisor	Safety attitude of immediate supervisor	3.67	4.63
	Control of immediate supervisor	3.93	3.84
	Conflict and performance management	3.54	3.49
Safety perception	Workload & fatigue	3.87	3.36

5 Discussion

During the development of the safety culture assessment tool, our first priority is to make certain the tool is reliable and valid. The analysis results showed us an acceptable reliability of the safety culture questionnaire in Table 3. With the continuing feedback from CAL's experts during the study, we used a series of expert panel to ensure questionnaire covers important issues of safety culture and to verify the content validity as well. In the further study, we would use confirmatory factor analysis to examine the construct validity of the framework in details.

Conducting a safety-related study is always sensitive and not easy in many industries. It is normal to only got hundreds of samples even the researchers have paid lots of efforts on samples collection. Without the full support of CAL in this study, it would be impracticable to have so many subjects (more than 1,000) in a single study to assure the representative meaning of the population. We believe this is a win-win situation and a success cooperation mode in aviation industry since the implementation of SMS. Both the industry and the academia need each other's support to ensure the safety culture evaluation is reliable and effective.

The quantitative data collection and analysis in this survey is just our first step. Before we could provide any practical recommendation to the airlines, we need further information regarding the hidden reasons and latent problems behind those low scored matters. These topics are worthy for further exploration to help the managers understand the inner thoughts of the subordinates. Therefore, the statistics results in this study will be used as discussion topics in our second phase, focus group, to help us not only understand the causes of lower scored issues such as departmental interaction but develop useful remedies of safety culture as well.

6 Conclusion

In this study, we combined both qualitative and quantitative methods to take the academic studies of safety culture into account and to meet the demands of the operation environment. Without the expert panel's opinions, it would be much harder for us to effectively develop a reliable and valid safety culture assessment tool. Although the statistical results show both the strengths and weaknesses of safety culture, the hidden reasons of the shortcomings or unstated ideas of workers are still remained undiscovered. Our next step is discussing the issues found in the study by focus group to further explore practical ideas of safety culture improvement.

References

1. Cole, K.S., Stevens-Adams, S.M., Wenner, C.A.: A Literature Review of Safety Culture. Sandia National Laboratories (2013)
2. Cooper, M.D.: Towards a model of safety culture. Safety Science 36, 111–136 (2000)

3. Cox, S., Cheyne, A., Alexander, A.: A Safety culture in offshore environments: developing the safety culture climate measurement tool. Paper Presented at the Proceedings of Safety Culture in the Energy Industries, University of Aberdeen (1997)
4. Geller, E.S.: Ten principles for achieving a Total Safety Culture. Professional Safety 39(9), 18–24 (1994)
5. HSE. A review of the safety culture and safety climate literature for the development of the safety culture inspection toolkit: Human Engineering (2005).
6. IAEA, The Chernobyl Accident: Updating of INSAG-1. Vienna: International Atomic Energy Agency (1992)
7. Liou, J.J., Yen, L., Tzeng, G.H.: Building an effective safety management system for airlines. Journal of Air Transport Management 14(1), 20–26 (2008)
8. McDonald, N., Corrigan, S., Daly, C., Cromie, S.: Safety management systems and safety culture in aircraft maintenance organisations. Safety Science 34(1), 151–176 (2000)

An Analysis of Hard Landing Incidents Based on Flight QAR Data

Lei Wang [1,2,3], Changxu Wu[1], Ruishan Sun [3], and Zhenxin Cui [3]

[1] Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Beijing, 100101, China
[2] Graduate University of Chinese Academy of Sciences, 19 A, Yuquan Road, Beijing, 100049, China
[3] Civil Aviation University of China, 2898 Jinbei Road, Tianjin, 300300, China
wanglei0564@hotmail.com,
changxu.wu@gmail.com

Abstract. Hard landing is one kind of typical landing incidents that can cause passenger discomfort, aircraft damage and even loss of life. This paper aimed to find out flight performance and operation features of hard landing incidents by using the methods of variance analysis, regression modeling and flare operation analysis based on flight QAR data. Results showed that pilots need to control the aircraft to an appropriate groundspeed and descent rate before descending to the flare initial point. Then control column and throttle operation in flare maneuver would affect landing performance conjointly. The logistic model showed that the vertical load of touching ground was actually linked with touchdown attitude and configuration closely, including three variables of pitch angle, roll angle and flap degree. These findings were expected to be applied in practice to prevent hard landing incidents and even landing accidents.

Keywords: Hard landing, QAR, flight safety, flare.

1 Introduction

Final approach and landing is the most important flight phase because a pilot needs to deal with more operations, decision-making, and workloads than other phases [1-4]. Hard landing is one kind of typical landing incidents which is defined as the main landing gear impacts the ground with a greater vertical speed and force than in a normal landing. Hard landings can vary in seriousness from simply causing mild passenger discomfort to situations resulting in serious vehicle damage, structural failure, and even loss of life [5-7]. When an aircraft has experienced a hard landing it has to be checked for damage before its next flight. Statistics also showed that hard landings happened frequently.

Though many studies regarding hard landing have been conducted, most of them have been based on models or experiments rather than real flight data [8-10]. Quick Access Recorder (QAR) is an airborne system which can record all kinds of position parameters, movement parameters, operation and control parameters, and alarm

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 398–406, 2014.

information in the whole flight phase. The hard landing was judged by the parameter of vertical acceleration. It is generally monitored by using QAR data in most commercial air carriers, but these data are also confidential for them. Meanwhile, there are few aviation administrators whom enforced their airlines to install QAR equipment on every commercial transport jet [13]. Therefore, QAR data were difficult and rarely utilized into research. This paper aims to find out flight performance and operation features of hard landing incidents through analyzing QAR data and put forward the prevention measures at the same time.

2 Methods

2.1 QAR Data Collection and Processing

The 119 cases of QAR data in this study were collected from three commercial aircrafts (Boeing 737-800) of a local airlines company. The original data is a CSV (Comma Separated Value) file with thousands of rows and columns. Therefore, VBA (Visual Basic for Applications) programing functions in Microsoft Excel was applied. In the final landing stage, aircrafts always fly within profile of a landing glide path; their position changes in the lateral axis are quite limited. Therefore, we focused on longitudinal and vertical parameters in this study. Finally, 19 columns of relevant original QAR data of every file were refined. 21 flight parameter variables were then selected and calculated as shown in the following table based on VBA programs. These parameter variables covered all flight and operational parameters in the critical visual and manual landing stages from the flare initial height to touchdown. Meanwhile due to flare maneuver would reduce the aircraft's descent rate to acceptable levels so that it settles gently on the main landing gear. It was seemed as one of the most skilled operation in flight [10], the pilot operation below 200 feet, especially the flare operation was selected as the main subject for analysis.

Among that the *Flare height* meant the height of initiating flare operation and *Flare time* meant the total time of aircraft flying from flare initial point to touch down point. It should be noted that the variable of flare time means the total time from flare initial point to touch down point. In addition, the flare operation initial point in this study is higher than the standard 30 feet in most flight manuals. This is because any slight backwards pulling of the control column could be recorded by a Quick Access Recorder, causing that the time and height of flare is earlier than theoretical value. The variable of *Touchdown Distance* and *Vertical Acceleration Touchdown* were two parameters using to determine long landing and hard landing. The *Vertical Acceleration Touchdown* meant the maximum value of vertical acceleration when the main landing gears touch the ground [11]. Based on the common statistical results of QAR data and monitoring criterion of aviation operators [12-14], the threshold of determining hard landing for this aircraft type was set as 1.4 g in this study.

Table 1. Selection of parameters

Classification	Name	Parameter name in QAR	Units
Kinematics & Performance	Flare height	RADIO HEIGHT	Feet
	Flare time	/	Second
	Groundspeed	GROUND SPEED	Knot
	Descent rate	VERT SPD	Feet/min
	Airspeed	AIR SPD	Knot
	Vertical acceleration	VERT ACCEL	g
	Touchdown distance	/	Feet
Operational Parameter	Throttle resolver angle	SELTD TRA FILTERED	Degree
	Control column position	CONTRL COLUMN POSN	Degree
	Control wheel position	CONTRL WHEEL POSN	Degree
	Control column force	CONTRL COLUMN FORCE	LBS
	Control wheel force	CONTRL WHEEL FORCE	LBS
	Flap handle position	FLAP HANDLE POSN	Degree
	Speed brake handle posi-	SPD BRAKE HANDLE	Degree
	Rudd pedal position	RUDD PEDAL POSN	Degree
Configuration & Attitude	Flap	FLAP	Degree
	Aileron	AILERON POSN	Degree
	Elevator	ELEV POSN	Degree
	Rudder	RUDD POSN	Degree
	Pitch angle	CAP DISP PITCH ATT	Degree
	Roll angle	CAP DISP ROLL ATT	Degree

2.2 Statistical Analyzing and Modeling

119 QAR data samples were divided into two groups with 65 cases of normal landing (Group 1) and the other one was 54 cases of hard landing (Group 2). QAR data of 65 normal landing events and 54 hard landings were regarded as two groups of independent samples. Each flight parameter variable of these 119 flights was also calculated by using VBA program.

First, the flight performance parameters of all flights were analyzed. For the aim of observing dynamic change of flight performance parameter variables in final landing phase and their differences between two groups, the altitude of 200 to 0 feet was divided into four flight phases (200-150-100-50-0 feet) and selected flight parameter was measured and compared in every phase. The multivariate analysis process of general linear model was introduced to compare the differences in the two groups.

Second, for the aim of finding the operation features of hard landing incidents and their correlations with landing performance, the statistical methods of variance analysis was used to find the difference of flare operation between normal landing and hard landing, including their parameter differences at flare initial point and in the whole flare process. One way ANOVA was used to examine variables which were subjected to normal distribution and non-parameter K-W test for other ones.

Third, aiming to find key flight parameters causing hard landing incidents, the logistic regression model on hard landing incidents was developed. In this study, the occurrence of hard landing was defined as a binary and dependent variable, where the value is 1 if it happened and 0 if it did not happen. The hard landing was judged by the parameter of vertical acceleration. We selected 17 flight parameters from Table 1 as original covariates in this logistical model, which including all operational parameters, configuration & attitude parameters and 3 kinematics parameters of groundspeed, airspeed and descent rate. Due to flare is a continuous operation from flare initial point to touchdown, the parameter value both at flare initial point and touchdown point were sampled in and there were 34 independent variables in total. The name and definition of each flight parameter is as showing in Table 1. The forward stepwise method was then performed. The likelihood ratio test (χ^2 difference) testing the change in –2LL (log likelihood) between steps was utilized to determine automatically which variables to add or drop from the model. The final predictor variables and coefficients of the model were obtained in the stepwise process. Simultaneously, the effectiveness of the model was checked and discussed below.

3 Results

3.1 Flight Performance Analysis

The variable of vertical acceleration is essentially both subjected to normal distribution and the results of Anderson-Daling test also proved it ($p > 0.05$). For the selected 119 samples, the mean and standard deviation of Vertical Acceleration Touchdown respectively was 1.387±0.082. The differences between 18 variables from 200 feet to touchdown were analyzed by using repeated measure and one-way ANOVA. Here only several important results regarding parameters of groundspeed, descent rate, control column and throttle are presented. Groundspeed and Descent rate are the two most flight performance parameters in landing and their change trend is as showing in Figure 1.

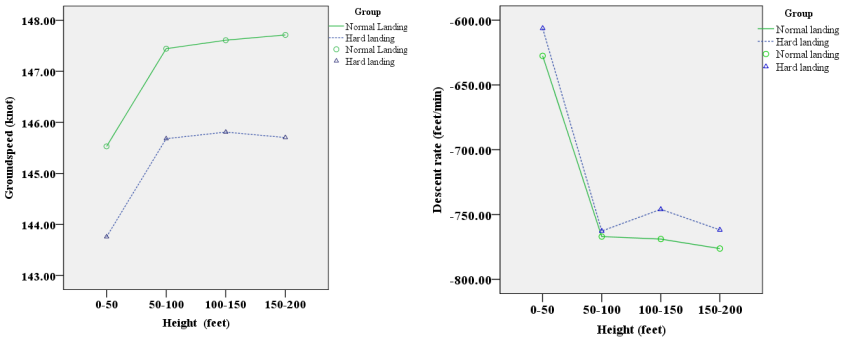


Fig. 1. Difference analysis of groundspeed and descent rate

As shown in Figure 1, the groundspeed of normal group is slightly greater than hard landing group. The difference of variable Groundspeed is not significant in the whole stage of 200-0 feet ($F(1,\ 117) = 1.763, p < 0.183$). The results of repeated measure ANOVA showed that the group effect of variable Descent rate is not significant ($F(1,\ 117) = 2.410,\ p = 0.123$). The descent rate of hard landing is slightly larger than the normal group before 50 feet, also the flare initial point, which changes a lot past 50 feet.

3.2 Flight Operation Analysis

Then, there were 65 normal landing samples (Group 1) and 54 hard landing samples (Group 2). The descriptive statistic on flare initial height and operation time of the two groups is as follows.

Table 2. Statistics on flare height and time

Group	**N**	**Flare Height** (*M* ± *SD, feet*)	**Flare Time** *(M* ± *SD, s)*
Normal Landing	65	52.169 ± 23.521	8.031 ± 2.076
Hard Landing	54	51.963 ± 20.175	7.722 ± 2.141

As seen in Table 2, there is no significant difference between the flare initial height of two groups, which are both around 50 feet ($F(1,\ 117) = 0.006, p = 0.941$). Meanwhile, the flare time of two groups also does not indicate significant difference ($F(1,\ 117) = 0.633, p=0.428$). Flare operation is considered one of the most technically demanding aspects of piloting. The results of the difference analysis on variables at the flare initial point are as shown in Table 3.

Table 3. Difference analysis on variables of flare initial point

Parameter Categories	Variable Names	Group	Mean±SD	*p*(ANOVA/K-W)
Operation Parameter	Throttle Resolver Angle	Normal	49.277±1.786	0.069
		Hard	49.922±2.044	
	Control Column	Normal	1.066±0.882	0.342
		Hard	0.930±0.616	
	Column Force	Normal	2.111±1.010	0.620
		Hard	2.024±0.860	
	Control Wheel	Normal	0.052±10.179	0.586
		Hard	0.954±7.233	
	Wheel Force	Normal	0.001±0.462	0.624
		Hard	-0.038±0.375	
	Flap Handle Position	Normal	30.462±2.115	0.000
		Hard	32.963±4.609	
	Speed Brake Position	Normal	2.991±0.888	0.540
		Hard	2.898±0.740	
	Rudder Pedal	Normal	0.570±0.314	0.747
		Hard	0.555±0.142	
Configuration and Attitude	Elevator	Normal	2.423±1.088	0.376
		Hard	2.576±0.720	
	Aileron	Normal	1.383±2.124	0.441
		Hard	1.650±1.500	
	Flap	Normal	30.462±2.115	0.000
		Hard	32.963±4.609	
	Rudder	Normal	-0.134±0.693	0.600
		Hard	-0.192±0.484	
	Pitch Angle	Normal	1.600±0.603	0.012
		Hard	1.301±0.677	
	Roll Angle	Normal	-0.245±1.437	0.327
		Hard	-0.466±0.894	
Flight Performance	Air Speed	Normal	148.923±4.748	0.259
		Hard	147.907±5.003	
	Groundspeed	Normal	145.815±7.104	0.461
		Hard	146.833±7.883	
	Descent Rate	Normal	-803.077±121.718	0.638
		Hard	-813.481±117.407	
	Vertical Acceleration	Normal	1.051±0.040	0.298
		Hard	1.044±0.033	

For the normal landing and hard landing groups, there are only three variables representing the significant difference at the level of 0.05, which are *Flap Handle Position, Flap* and *Pitch Angle*.

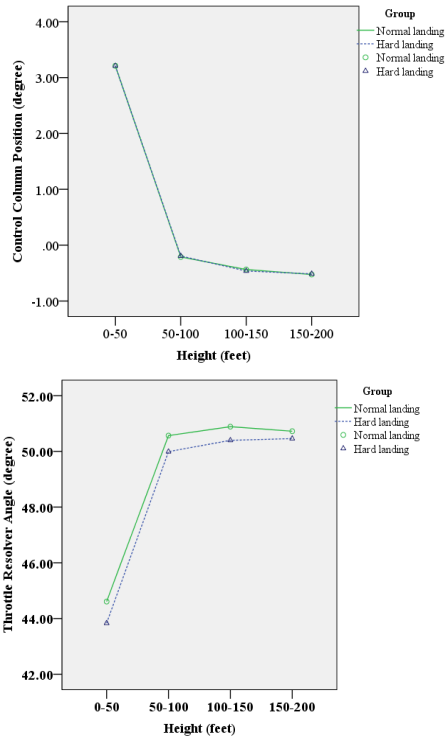


Fig. 2. Difference analysis of control column and throttle resolver angle

In Figure 2, the control column and throttle change greatly after passing 50 feet (flare operation initial point). There is no difference between the control column of the two groups ($F(1,\ 117) = 0.000, p = 0.998$). There is also no difference found for throttle operation before 50 feet ($F(1,\ 117) = 3.349, p < 0.07$). The difference is reflected after a flare starting when the pilot begins to decrease thrust.

3.3 Logistic Regression Model on Hard Landing

Table 4 shows the estimated parameters of the logistic model in predicting landing incident type (hard landing or normal landing). Three predictors were included in the final logistic regression model. The overall predictive percentage of the model was 72.6%, the sensitivity was 0.697and the specificity was 0.786.

As shown in Table 4, the Wald criteria indicated that *Flap Handle Touchdown, Pitch Angle Touchdown* and *Roll Angle Touchdown* significantly contributed to the occurrence of hard landings ($p < 0.01$). Nagelkerke's R^2 of 0.677 indicated a relatively strong relationship between predicting variables and hard landing.

Table 4. Logistic regression values of the predicting variables

Predicting variables	Wald (χ^2)	Adjust OR [a]	95% C.I.for OR [b]
Flap Handle Touchdown	11.107**	1.172	1.074-1.296
Pitch Angle Touchdown	18.613**	0.531	0.393-0.713
Roll Angle Touchdown	15.984**	2.229	1.489-3.281
Constant	3.469#	0.058	

** $p < .01$, * $p < .05$,# $.05 < p < .10$ and otherwise $p \geq .10$.

[a]Adjust ORs (odds ratio) predicted hard landing.

[b]Confidence interval.

4 Discussion and Conclusions

In aviation safety research, the focus has typically been more on aviation accidents where their occurrence rate has been decreased to quite a low level in most regions of the world. However, unsafe incidents have often been ignored due to the difficulty in obtaining and analyzing them in detail. Basing on flight QAR data, this study provided a new way to analyze unsafe incidents in the landing phase by considering a history of individual instances recorded during flight. The main findings in this study were concluded as following.

1. The results of multivariate analysis indicated that most flight parameter variables with differences appeared in the stage of 50 feet to touchdown. Theoretically speaking, many flight operations, including flares, need to be finished by pilots in just a few seconds. While aircraft in low speed flight is sensitive to wind and other weather factors, any small configuration changes during this stage could easily complicate the decision of the proper action to take at the decision point. Therefore, this phase is the most important operation stage and pilots should check the ratio of descent rate and groundspeed carefully at the point of 50 feet.
2. Flare would reduce the aircraft's descent rate to acceptable levels so that it settles gently on the main landing gear, it would greatly influence vertical acceleration through the two key factors of flare time and final flare pitch angle. The control column and throttle operation would affect landing performance conjointly. Pilots' quick and steady pulling up columns and softer throttle reduction are helpful for a better flare operation and better landing performance.
3. The logistic model showed that the vertical load of touching ground was actually linked with touchdown attitude and configuration closely, including three variables of pitch angle, roll angle and flap degree. Among these, the pitch angle of the aircraft is correlated with control column operation directly and therefore is a main external indication of flare. As a matter of fact, the correlations between pitch angle and vertical acceleration were strong at every stage from 200 to 0 feet.
4. These findings would be the basis of developing a mathematical and quantitative model for further revealing the relationships between pilot operation and landing

performance, which can also be applied in practice to prevent hard landing incidents and even landing accidents.

Acknowledgments. We appreciate the support of this work from the National Natural Science Foundation of China (No. 61304207, No.U1333112), the National Basic Research Program of China (No.2010CB734105) and the Fundamental Research Funds for the Central Universities (No. ZXH2012D001).

References

1. Hawkins, F.H.: Human Factors in Flight, 2nd edn. Ashgate. Brookfield, VT (1993)
2. Wickens, C.D., Hollands, J.G.: Engineering Psychology and Human Performance, 3rd edn. Prentice Hall Press, Upper Saddle River (2000)
3. Harris, D.: The influence of human factors on operational efficiency. Aircraft Engineering and Aerospace Technology 78(1), 20–25 (2006)
4. Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., Wiegmann, D.: Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. Human Factors 49(2), 227–242 (2007)
5. Sartor, P., Bond, D.A., Staszewski, W.J., Schmidt, R.K.: Value of an overload indication system assessed through analysis of aviation occurrences. Journal of Aircraft 46(5), 1692–1705 (2009)
6. Nie, L., Shu, P., Huang, S., Wang, X.: Intelligent Diagnosis for Hard Landing of Aircraft Based on SVM. China Safety Science Journal 19(7), 149–153 (2009)
7. Wang, X., Shu, P., Cao, L.: Incremental AHP based Weight Assessment for Risk Factors of Civil Aircraft Landing Process. China Safety Science Journal 20(9), 112–115 (2010)
8. Rosa, M.A.V., Fernando, G.C., Gordún, L.M., Nieto, F.J.S.: The development of probabilistic models to estimate accident risk (due to runway overrun and landing undershoot) applicable to the design and construction of runway safety areas. Safety Science 49(5), 633–650 (2011)
9. Molesworth, B., Wiggins, M.W., O'Hare, D.: Improving pilots' risk assessment skills in low-flying operations: The role of feedback and experience. Accident Analysis and Prevention 38(5), 954–960 (2006)
10. Benbassat, D., Abramson, C.I.: Landing flare accident reports and pilot perception analysis. International Journal of Aviation Psychology 12(2), 137–152 (2002)
11. Civil Aviation Administration of China. Implementation and Management of Flight Operation Quality Assurance. Advisory Circular: 121/135-FS-2012-45, CAAC, Beijing, China (2012)
12. Qi, M., Shao, X., Chi, H.: Flight operations risk diagnosis method on Quick Access Record exceedance. Journal of Beijing University of Aeronautics and Astronautics 37(10), 1207–1211 (2011)
13. Wang, L., Wu, C., Sun, R.: Pilot operating characteristics analysis of long landing based on flight QAR data. In: Harris, D. (ed.) EPCE 2013, Part II. LNCS, vol. 8020, pp. 157–166. Springer, Heidelberg (2013)
14. Sun, R., Han, W.: Analysis on parameters characteristics of flight exceedance events based on distinction test. China Journal of Safety Science and Technology 7(2), 22–27 (2011)

Study on Eye Movements of Information Omission/Misjudgment in Radar Situation-Interface

Xiaoli Wu[1,2], Chengqi Xue[1], Yafeng Niu[1], and Wencheng Tang[1]

[1] School of Mechanical Engineering, Southeast University, Nanjing 211189, China
{ipd_xcq,niuyafeng}@seu.edu.cn, fanyn09@163.com
[2] College of Mechanical and Electrical Engineering, Hohai University, Changzhou 213022, China
wuxlhhu@163.com

Abstract. Radar situation interface belongs to a sub-interface of a complex system. Because the information in human-computer interaction interface of a complex system is of a large amount and in complicated relationships, it is apt to cause misreading, misjudgment and information omission in the target search. The critical factors causing error problems like information omission and misjudgment in the radar situation interface are analyzed. Based on the behavioral data and the physiological data derived from eye movement tracking, the misperception factors leading to users' information omission/misjudgment are detected. The experimental results showed that, (1) Both interval size and vision position impose a significant influence on the visual cognition of target search. The interval should not be too large for target search in the situation interface, otherwise it may result in long reaction time and omission and misjudgment. (2) During the target search in the upper vision, lower vision and peripheral vision, the reaction time and the error rate present significant changes, and the reaction time of peripheral vision achieves the longest. The vision position also exerts a remarkable influence on the first saccade latency. The fixation duration and fixation point number display obvious changes, and the mean fixation duration of the lower vision is the longest while its fixation point number is the smallest, which is apt to cause misjudgment and omission of information. (3) Eye movement plots can effectively reflect the process of information search, and the gaze plot and the heat point map can present the relevant factors of information omission. And the conclusion reached can be used as reference for the information design and layout of the situation interface of future complex system, so as to effectively improve the misperception problems like omission and misjudgment in the target search process.

Keywords: Radar Situation Interface, Information identification, Omission, Misjudgment, Misperception, Visual perception, Eye movements.

1 Introduction

With the rapid development of industrial design and computer interactive media, visual information interface has become an essential information interactive medium in a

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 407–418, 2014.

complex system. The unreasonable design of interface information has given rise to malfunctions of cognition and decision-making among operators, thus leading users into a complex cognition and finally resulting in serious failures in information recognition and analysis, and even in operation and execution processes, which poses one of the major causes for many accidents. Errors are common human failures occurring in information interface and its cognition mechanism of errors is an important hitting-point for improving interface design as well as the key for reducing cognition difficulties. Radar situation interface belongs to a sub-interface of a complex system. Because the information in human-computer interaction interface of a complex system is of a large amount and in complicated relationships, it is apt to cause misreading, misjudgment and information omission in the target search. It exerts a significant influence on improving the interface layout for effectively detecting misperception factors causing information omission, misreading and misjudgment.

2 Background

Error problems like information omission/misjudgment are mainly originated from the studies on error factors. The studies conducted by domestic and foreign researchers on systematic disorder and human error mainly focus on error analysis and human reliable methods. Embrey, Altman, and Swain, et al. tried to use the basic behaviour component of the operator to describe the behavior of the operator with "error" event characteristics from the view of traditional human factors; PHEA and HRMS et al. established the analysis model of human factor from the perspective of cognitive psychology. Nielsen[1] (1994) and Shryane[2] (1998) proposed the availability interface design method to reduce human error probability (HEP). Hidekazu[3] (1999) studied human error probability (HEP) through user evaluation model. And Krokos and Baker[4] (2007) also proposed interface cognition error classification method. Maxion[5] (2005) improved operation interface dependability through mitigation of human error (External Subgoal Support). Li Pengcheng[6], (2011) conducted a study on human error and reliability in digital control system of nuclear power plant, analyzing the error model of the missions of operators in nuclear power plant and establishing the risk evaluation model of human factor reliability of digital control system. Shappell[7-8] (2001, 2007) demonstrated the corresponding cognitive factors, like attention and understanding, with the technology and decision after analyzing the aircraft accident data from past 13 years; concluded the error probability in perception level was relatively low.

Domestic and foreign studies on the complex information interface mainly emphasis on the reasonability evaluation method of interface design; Wilson[9] (2005) used eye tracking technology to conduct the trial on FS35 fighter interface, tested the prospective memory and attention diversion of pilots and determined the cognitive complex factors of pilots. Wang Haiyan, Xue Chengqi[10] (2011) have experimentally evaluated and analyzed the layout design of fighter radar situation-interface through an objective evaluation technology of eye tracker, and selected a rational special layout optimization scheme through the evaluation by eye moving data indexes. Liu

Qing [11] (2012) simulated the general operation sequence of enemy attack task in avionics system to conduct interface design on infrared radar system, navigation system, weapon mounting system and flight control system, and evaluated the rationality of interface layout, navigation and graphic symbols by the eye tracker experiment. Li Jing and Xue Chengqi [12]et al. (2012) have studied the influence of the time pressure of complex digital interfaces on color and shape codes, to explore the identification performances under different time pressures. Dong Xiaolu[13] (2010) researched time press impact of digital interface from human error. For all the studies mentioned above, the experimental data are adopted as the main evaluation method to judge the rationality of complex information interface design. Few scholars have set foot in such fields like the reasons of information omission/misjudgment.

3 Method

The paper simulated the radar situation-interface of complex system. The nested cognitive experiment of reaction time and eye movement tracking was conducted. The analysis of variance method was applied to statistically analyze the indexes of reaction time and error rate as well as the indexes of gaze plot and fixation duration in eye movement data and the issues were discussed how the visual position and interval influenced target search and resulted in problems like omission and misjudgment caused by misperception. The experiment was divided into two parts: first, E-Prime was adopted to design software for reaction time experiment on stimulus features and then physiological measurement technology was employed for eye moving experiment on visual position. The study mainly analyzed the influence of eye movement indexes including total fixation duration, fixation count and saccade latency on the error factors of information omission/misjudgment.

4 Experiment 1: Visual Confined Research

4.1 Method

The experiment discussed the visual search in radar situation interface from two variables of visual position and target object interval, and founded that the visual position and target object interval served as the key factors which affected the visual limitation to result in omission and misjudgment. According to the six feature items of samples, the experiment will investigate the numbers of enemy plans, friend planes and unidentified objects that will appear. Experiment 1 adopted 3×2×3 with-in group design, with the three factors respectively being quantity of target objects (2, 4 and 6), visual position (upper vision and peripheral vision) and target object interval (12mm, 48mm and 96mm). The experimental procedure was written by the professional psychological experiment development software E-Prime. The experiment was conducted in the Human-Computer Interaction Lab of Hohai University, and the experimental subjects were 20 undergraduate students in the university.

4.2 Result

The study adopted the simulated fighter information matters as the experimental materials and conducted the experiment from the point of errors. The experimental result showed that, during information identification, with the gradual increase amount of information matters (two, four and six), the reaction time of target objects in peripheral vision display a gradually increasing trend compared with those in upper vision, and that the reactions of searches for target objects with different intervals reached remarkable levels. When enemy planes, friend planes and unidentified objects appeared in different visual positions of attack interface in the form of simulants and were presented in different numbers and intervals, the reaction times and error rates of subjects were shown in Fig.1. The variance analysis on reaction times showed that, the main effect of intervals of upper visual positions ($F=14.416$, $P=0.012$, $p<0.05$) and that of peripheral visual positions ($F= 6.990$, $P= 0.00103$, $p<0.05$) both reached remarkable levels. The variance analysis on error rates showed that, the main effect of intervals of upper visual positions ($F=2.380$, $P=0.013$, $p<0.05$) and that of peripheral visual positions ($F=9.308$, $P= 0.014$, $p<0.05$) reached remarkable levels. Hence, the size of intervals can exert a significant influence on the visual cognition of target search in visual positions. The analysis extracted the data collected under a high error rate and found that, the target search with different interval information showed a linear increase on the error rate, while displaying an inverted-V during the reaction. Therefore, errors like misjudgment and omission did not necessarily require the longest reaction time.

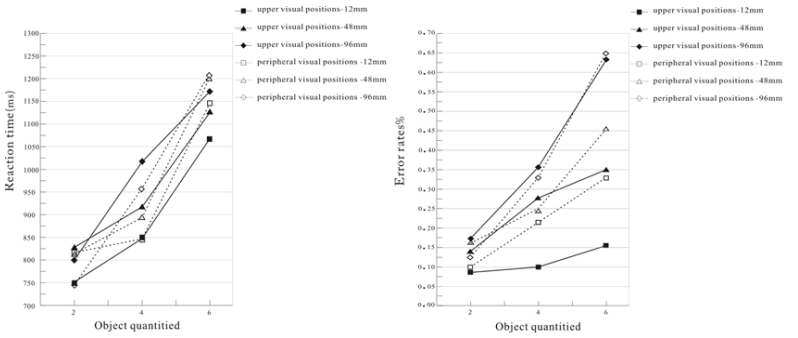


Fig. 1. Reaction time and Error rates in three intervals under different vision positions

5 Experiment 2: Eyes Movement Research

5.1 Method

Material. The stimulus feature items were designed as the simulated information matters in radar AD-attack situation interface of a complex system, specifically, they were the enemy planes, friend planes and unidentified objects, including six objects of green friend planes, red enemy planes and four unidentified objects of green, red and

yellow) whose size was fully filled up in black boxes. The enemy planes, friend planes, red and green unidentified objects appeared alternatively. Since red and green colors existed in the attack situation interface which possessed a strong interference, the yellow unidentified objects only serves as interference items rather than the target objects for investigation to reach the same interference strength of target objects. The stimulus feature items were designed as the same in radar AD-attack situation interface in experiment 1. All the materials are presented in the attack situation interface with a radius of 78mm. The circle with a radius of 39mm was the attack range and the pattern of the host computer was in the center. Various data of current situation were presented in around mainly in white, red and green, which would further impose fixed intervention on subjects. All of these presentation elements were quantitative data and formed a complex situation environment. However, the enemy planes, friend planes and unidentified objects acted as variables and were presented within the annulus constituted by large circles and small circles.

Apparatus. The experiment was conducted in the eye movement tracking laboratory of HHU (Hohai University). The Switzerland-made tobii1X120 eye tracker with a sample frequency of 120HZ and gaze location precision of 0.5 degrees was adopted. The computer with a display pixels of 1280×1024 (px), a color quality of 32-bit, a collection way of eyes collection and a head movement range of 30×16×20cm, was was adopted. The sight-line gaze location data of the system were delayed to 3ms and possessed an ideal gaze - instantaneous display. The system took samples from the eyeballs of subjects every 20ms, to investigate and collect the data of eyeball movement of subjects.

Participants. The experiment was conducted in the Human-Computer Interaction Lab of Hohai University, and the experimental subjects were 20 undergraduate students in the university, 10 females and 10 males, aging between 19 and 23 years old, with normal vision or corrected vision, and without color blindness or color weakness. Before the experiment, relevant information of the subjects were input, including their gender, age, major and vision.

Procedure. The subjects were first required to be familiar with the task environment of radar attack situation; information in such situation environment remained unchanged and the interference from the information should be avoided. Then, the features (colors and shapes) of target objects including enemy planes, friend planes and unidentified objects were informed, which would be considered as the objects of target search. At the beginning of the experiment, the subjects were asked for the number of enemy planes, friend planes or unidentified objects appearing in the situation environment and required to remember the targets they were searching for; the presentation of the target objects was set at 1,500ms. After the screen was blackened, nine stimulus items would appear, among which, 6±1 target objects were required to be searched by subjects; during the presentation of 1,500ms, the subjects were asked to rapidly figure out the number of target objects appeared and press the number key for reaction after the black screen. It cost about 15 minutes for each one to perform an entire experiment.

Design. On the basis of analysis the data from experiment 1, experiment 2 adjusted the setting of independent variables and further discussed the error-prone (omission and misjudgment) visual limitations. Six target objects were regarded as the basic quantity of visual search and then were adjusted to 6±1 in order to cause quantitive randomness to subjects. To further understand the positional characteristics, the lower vision factor was added. The interval design had excluded the single interval of 12mm and mainly probed into four fold and eightfold intervals.

5.2 Result and Discussion

Gaze Plot. The data derived at the earliest by eye movement tracking experiment is the Gaze Plot. The gaze plots in different vision locations were shown in Fig.2. At the field of upper vision, the subjects could focus on searching the number of enemy planes in the upper visual area; when subjects searched the enemy planes in the lower visual area, it could be clearly detected that, the subjects would spontaneously glance the upper visual area and be interfered by the upper vision. However, their gaze plots in peripheral visions were more scattered and the integral glancing paths covered a larger area of upper vision. As a result, a discussion on the information search in radar situation interface based on vision locations is of distinctive features. The factors causing information omission and misjudgment can be discussed from the perspectives of the gaze plot and the location of target objects.

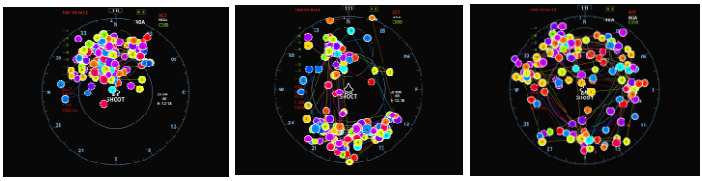


Fig. 2. Fig.2 Gaze plots under different vision positions (Left: the upper vision; Middle: the lower vision; Right: the peripheral vision)

The subjects were asked to figure out the total number of the enemy planes appeared. Fig.3 showed the gaze plots of one subject's target search in the upper vision and the lower vision, respectively. The left figure displayed seven fixation points of the subject; the first fixation point is near the center and then moved to and stayed at the red data part; afterwards, the fixation point continued to stay from the influence of icons in the upper attack range; the subject began to search the target objects again until the sight reached the fourth fixation point. Obviously, it was difficult for this subject to search out all enemy planes within 1,500 ms, and which lead to the omission of target objects. The number of fixation points of the subject in the right figure was eight, but his former three fixation points all stayed at the red area within the attack range and was transferred to the target objects of lower vision until reaching the fourth fixation point. This demonstrated during target search, other irrelevant non-target information matters (the unidentified objects and information data display, etc.) become the interferents of information search and occupy the fixation duration, therefore, causing attention shift.

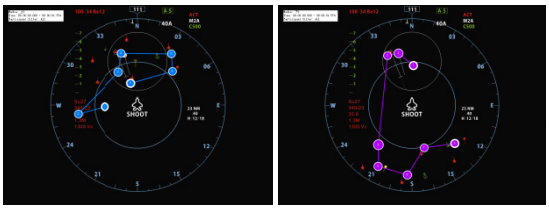


Fig. 3. Gaze plots of one subject's target search respectively in the upper vision and the lower vision

Heat Map. The heat map composed through the region of interests of twenty subjects shows the heat points distribution of the upper vision (the left figure in Fig.4), the lower vision (the middle figure in Fig.4) and the peripheral vision (the right figure in Fig.4). Among which, the brightest red area indicates the subject presents a longest gazing time and that the largest number of the fixation points distributed in that area. However, the heat point area simultaneously appearing both in the upper vision and the peripheral vision acts as the indicator icons of attack range, which suggests that this information indication has seriously impacted the target search and is the essential factor causing information omission. It can be further illustrated based on the duration and times of fixation points.

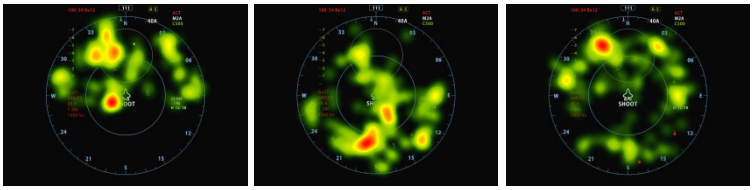


Fig. 4. The heat point map in different vision positions (Left: the upper vision; Middle: the lower vision; Right: the peripheral vision)

Total Reaction Time and Error Rate. When the enemy planes appear in the upper-middle, lower and peripheral vision positions of attack situation interface in the form of stimulants with other unidentified objects distributed around, the reaction time and error rate of the subject when the objects are presented in two different intervals are shown in Fig.5 and Fig.6. The variance analysis on the reaction time shows that, the interval main effects of the vision position（$F=81.227$，$P=0.004$，$p<0.05$）has reached remarkable levels. The variance analysis on the error rate suggests that, the interval main effects of the vision position ($F=7.562$，$P=0.021$，$p<0.05$) has reached remarkable levels. Therefore, the vision position exerts a significant influence on the visual cognition of target search.

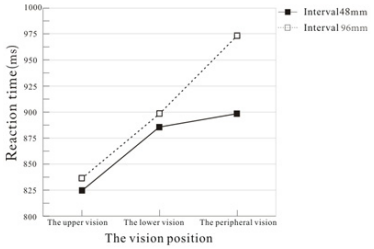


Fig. 5. The reaction time in different vision positions

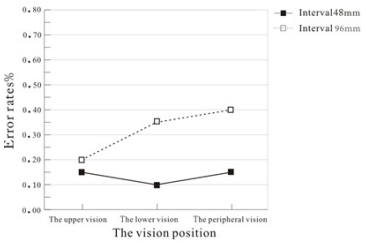


Fig. 6. Error rates in different vision positions

Mean of Fixation Duration. The mean fixation duration of fixation points refers to the ratio of the Total Fixation Duration to the Fixation Count in the task. During a normal visual observation, the performance of eye movement is reflected by the rapid saccade between a series of sight stay and the staying points on the observed object. The stay of eye movement for at least 100ms is generally called the fixation. Most information can be processed only during the fixation. Therefore, the process of target search is significantly influenced by the vision positions. The fixation duration and fixation times of the upper vision are greatly larger than those of the lower and peripheral vision. In addition, they can easily capture the targets easier with few information omissions, as it shown in Tab.1.

Table 1. Mean fixation duration and fixation times

		Fixation Duration（Mean）	Fixation Duration（N）
The upper vision	Interval（48mm）	240	5.80
	Interval（96mm）	285	6.55
The lower vision	Interval（48mm）	490	4.15
	Interval（96mm）	365	5.10
Peripheral vision	Interval（48mm）	235	6.15
	Interval（96mm）	215	6.65

Time to first fixation refers to the position where firstly gazed by the eyeballs of the subject. The information configuration of the situation interfaces at three different vision positions is shown in Fig.7. The possible visual search areas are divided according to the positions of target objects and interferents depending on the distribution condition of the heat point map. The visual search areas in eye movement tracking areas are called the areas of interest. In which, AOI 1-4 are the areas which can be easily gazed by the eyeballs and also is the area where target enemy planes are located. The mean fixation duration in different areas can be acquired. From the areas of interest in Fig.10 which corresponds with the data listed in Tab.2, AOI3 in the upper vision displays the longest mean fixation duration (760ms), which further explains that the digital information in the upper left corner has received the highest intensified interference and can easily cause attention shift. When target objects

appear in the lower vision, the time to the first fixation stays at the attack information indicating area (1040ms), followed by target objects presented in the lower vision (540ms). Since the color of the interferents (appearing in red) is the same as that of the target objects, it will easily lead to attention shift. As a result, the subject will easily take the unidentified objects as the enemy planes by misjudgment and lead to a wrong judgment.

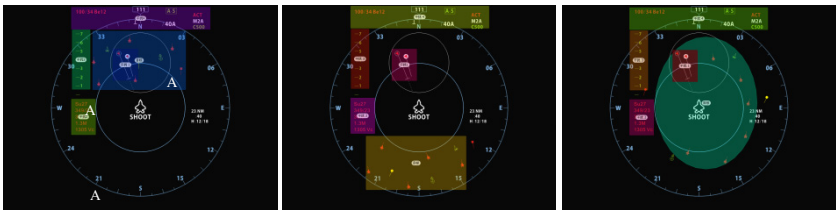


Fig. 7. AOS division at different vision positions (Left: the upper vision; Middle: the lower vision; Right: the peripheral vision)

Table 2. Fixation duration and fixation points divided by different areas of interests

Division of areas of interest	AOI 1		AOI 2		AOI 3		AOI 4		AOI 5	
	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N
The upper vision	330	19	170	1	760	1	-	-	150	20
The lower vision	1040	3	440	2	-	-	280	1	540	20
Peripheral vision	370	7	700	3	840	2	-	-	100	20

Saccade Latency. The first saccade latency refers to the time interval from the start of the stimulus presentation to the first saccade made by the subject, including the time required for information-in and out as well as the central processing time controlled by the pathways and its complexity. The variance analysis on repeated measure reveals that, the vision positions of target objects impose an appreciable influence on the first saccade latency of the subject. As it is shown in Fig.8, the reaction times in the vision positions have reached remarkable levels ($F= 3.571$, $P=0.036$, $p<0.05$). When the target object appears in the lower vision position, the first saccade latency of the subject is greatly longer than the presentation time of the target objects in the upper vision, which explains that the vision positions of the target objects have influenced the attention of the subject to some extent and lead to a delay in saccade time. The target object intervals exert an influence on the first saccade latency; relevant comparison shows that, the first saccade latency of the stimulus presentation in eight-time intervals is longer than that of the stimulus presentation in four-time intervals. The interactions between each variable are not obvious.

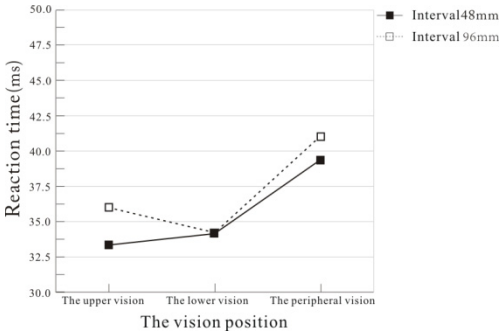


Fig. 8. Saccade latency

6 General Discussion

This experiment discusses the visual search of the radar situation interface based on two variables, i.e., vision position and target object interval. The experiment finds that, the vision position and target interval serve as the key factors influencing the visual limitation and further lead to errors like omission and misjudgment. The result of experiment 1 demonstrates that, during the information identification, with the increase amount of the information matters (two, four and six), the reaction time required by the target objects in peripheral vision is gradually increasing than that in the upper vision, and the target searches in different intervals have reached remarkable levels; the analysis extracts the data under high error rate and reveals that, during the visual searches of information with different intervals, the error rate displays a liner increase, while the reaction time presents an inverted-V change. Therefore, the errors like misjudgment and omission do not necessarily require the longest reaction time. Wickens et al. [15](1990) studied the identification on the information with different color codes and spatial positions under multiple information channels, and proposed the different influences of colors and positions on the information identification; Dukic and Hanson[16-17] (2005,2006) studied the visual search mode of pilots and proposed relevant strategies for visual search. Based on previous conclusions, experiment 2 selects the vision position and the target interval as variables. Since these two variables are the key factors of relevant visual limitation, and the target objects are easier to be omitted in the lower vision and peripheral vision. The result of experiment 2 indicates that, the vision position and the other information matter in visual area are the main factors that prone to cause the omission and misjudgment in visual search; particularly, in the lower vision and the peripheral vision, target objects are easier to be subject to omission. The eye movement study in experiment 2 further discusses the visual limitation prone to errors (omission and misjudgment, etc.) basing on experiment 1. Six target objects with high error rates are regarded as the basic amount of the visual search, and the experiment obtains the eye movement plot and heat map based on the data of eye movement experiment from three different visions. In addition, the fixation duration and fixation times of omission and misjudgment are also analyzed, and the rules of eye movement of visual search of subjects are concluded from

physiological data. The experiment finds that, because the eyes are accustomed to following the visual rule of from the left to the right and from the upper to the lower, during the target search in the lower vision, the subject will always start the search the other information matters from the upper vision, and begin the search for the target objects in the lower vision in the second and third saccade. The first saccade latencies in the lower vision and peripheral vision are obviously longer than those required by the target object presentation in the upper vision. Furthermore, the experiment finds that, the visual interference will also cause omission and misjudgment; sequence experiments on the visual interference factors are demanded. There are many information matters in the radar situation interface which are deserved to be simulated and tested in the classified simulated situation environment, which is meaningful for the further study.

7 Conclusion

Both interval size and vision position impose a significant influence on the visual cognition of target search. The interval should not be too large for target search in the situation interface, otherwise it may result in long reaction time and omission and misjudgment. During the target searches in upper vision, lower vision and peripheral vision, the reaction time and error rate present significant changes and the reaction time required in peripheral vision is the longest. The fixation time and fixation point number also show significant changes; the mean fixation time of lower vision is the longest, but the fixation point number is the smallest, which is easier to cause information misjudgment and omission. The eye movement plot can effectively reflect the process of information search; the fixation plot and heat map can present relevant factors of information omission; different vision positions and intervals also exert an appreciable influence on the first saccade latency. The vision position is not regarded as the most essential error factor; the data of eye movement explain the features of information matters, such as the color and shape, are easier to cause attention shift leading to information omission.

Application. The study was not conducted in a real fighter cabin and the information matters of enemy plane, friend plane and unidentified objects were all simulated, so the data obtained are simulated data, and the conclusion reached can be used as reference for the information design and layout of the situation interface of future complex system, so as to effectively improve the misperception problems like omission and misjudgment in the target search process.

Acknowledgment. This work was supported by the National Nature Science Foundation of China（Grant No.71071032,71271053，the Social Science Fund for Young Scholar of the Ministry of Education of China(Grant No. 12YJC760092), and Fundamental Research Funds for the Central Universities of China (Grant No. 2013B10214)，

References

1. Nielsen, J., Mack, R.L.: Usability Inspection Methods. Wiley, New York (1994)
2. Shryane, N.M., Westerman, S.J., Crawshaw, C.M., Hockey, G.R.J., Sauer, J.: Task analysis for the investigation of human error in safety critical software design: a convergent methods approach. Ergonomics 41(11), 1719–1736 (1998)
3. Yoshikawa, H., Wu, W.: An experimental study on estimating human error probability (HEP) parameters for PSA/HRA by using human model simulation. Ergonomics 42(11), 1588–1595 (1999)
4. Krokos, K.J., Baker, D.P.: Preface to the special section on classifying and understanding human error. Human Factors 49(2), 175–176 (2007)
5. Maxion, R.A., Reeder, R.: Improving user-interface dependability through mitigation of human error. International Journal of Human-Computer Studies 63(1-2), 25–50 (2005)
6. Pengcheng, L.: Study on human error and reliability in digital control system of nuclear power plant. South China University of Technology (2011)
7. Shappell, S.A., Wiegmann, D.A.: Applying Reason: The human factors analysis and classification system (HFACS). Human Factors and Aerospace Safety 1(1), 59–86 (2001)
8. Shappell, S., Detwiler, C., Holcomb, K.: Human Error and Commercial Aviation Accidents: An Analysis Using the Human Factors Analysis and Classification System. Human Factors 49(2), 227–242 (2007)
9. Wilson, J.R., Hooey, B.L., Foyle, D.C.: Head-Up Display Symbology for Surface Operations: Eye Tracking Analysis of Command-guidance vs. Situation-guidance Formats. In: Proceedings of the 13th International Symposium on Aviation Psychology, Oklahoma City, pp. 13–18 (2005)
10. Haiyan, W., Ting, B., Chengqi, X.: Experimental evaluation of fighter's interface layout based on eye tracking. Electro-Mechanical Engineering 27(6), 50–53 (2011)
11. Haiyan, W., Ting, B., Chengqi, X.: Layout design of display interface for a new generation fighter. Electro-Mechanical Engineering 27(4), 57–61 (2011) (in Chinese)
12. Qing, L.: Research on design and experiment method of avionics system display interface. School of Mechanical Engineering, Southeast University, Nanjing (2012) (in Chinese)
13. Jing, L., Chengqi, X., Haiyan, W.: Information encoding in human-computer interface on the equilibrium of time pressure. Journal of Computer Aided Design & Computer Graphics 25(7) (2013) (in Chinese)
14. Xiaolu, D.: Influence of human-system interface design method and time pressure on human error. Industrial Engineering Department of Tsinghua University, Beijing (2010) (in Chinese)
15. Wickens, C.D., Andre, A.D.: Proximity compatibility and information display: effects of color, space, and object display on information integration. Human Factors 32(1), 61–77 (1990)
16. Dukic, T., Hanson, L., Holmqvist, K., Wartenberg, C.: Effect of button location on driver's visual behaviour and safety perception. Ergonomics 48(4), 399–410 (2005)
17. Dukic, T., Hanson, L., Falkmer, T.: Effect of drivers'age and push button locations on visual time off road, steering wheel deviation and safety perception. Ergonomics 49(1), 78–92 (2006)

Analysis on Eye Movement Indexes Based on Simulated Flight Task

Chengjia Yang[1], Zhongqi Liu[2,*], Qianxiang Zhou[2], Fang Xie[3], and Shihua Zhou[1]

[1] Astronaut Center of China, Beijing 100094, China
[2] School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China
[3] General Technology Department, China North Vehicle Research Institute Beijing, 100072, China
liuzhongqi@buaa.edu.cn

Abstract. To probe pilot's attention allocation, workload and cognition by eye movement indexes analysis. Six subjects participated the experiment. They were asked to fly three simulation scenarios: landing, climbing and cruise flight. Five eye movement indexes which they were the percentage of fixation point, the percentage of dwell time, average fixation duration, average pupil size and average saccade amplitude were recorded and analyzed. The result indicated that eye movement data was obviously different in out view and instrument panel; also it was different among three tasks. Conclusions can be made from the result: subjects spent most time on outside view while leaving small time to make quick crosschecking to the instrument; Subjects show different pattern of attention allocation through three flight tasks; the recorded eye movement indexes are the good indicators to pilots' attention allocation, workload and cognition.

Keywords: Eye movement, Flight simulator, Attention, Workload, Cognition.

1 Introduction

It is well known that aircraft driving is a highly visual task during which pilots must fixate objects finely to finish accurate control and scan the environment quickly to acquire the situation awareness of flying aircraft. Eye movements indexes are the indicator of human mental activity. Pilots' attention allocation, the change of their workload and fatigue will all be clear by analyzing eye movements. The cockpit's design and instrument layout can be reached from eye movements' analysis. Consequential result is the best man machine interaction and the alleviation of pilot's workload. Fitts, etc. conducted a series of researches to pilots scanning behavior at the end of 1940s, and the result was the establishment of typical "T" of cockpit's instrument layout[1]. After Fitts' work, more and more researchers have researched on pilot's eye movement and eye tracking has been widely applied in aviation field. Eye tracking can be used to[2]: (1) compare eye scanning behavior of pilots using conventional displays and new display; (2) do behavioral assessment of pilot visual attention under

* Corresponding author.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 419–427, 2014.

various levels of visibility;(3) measure performance based on fixation duration and number of fixations;(4) evaluate pilot workload;(5) evaluate situation awareness (6) measure attention and fatigue, et al. To meet operational needs over the next 20 to 30 years, American Air Force introduced a revolutionary virtual crew station concept titled the "Super Cockpit"[3]. The pilot can interacts with the display spatially by pointing his eyes at objects in the display and giving verbal commands. Functions can also be activated by merely looking at a displayed switch and saying "select," or "on, or "off," or "go there," or "stop here," etc.

The purpose of this study is to explore the characteristics of pilots' attention allocation, workload change, and cognition by eye tracking of pilots' scanning behavior and the analysis of pilots' eye movement indexes. The work will provide some valuable reference to the design of aircraft cockpit.

2 Method

2.1 Subjects

Six young males participated the experiment. They have been trained proficiently enough to implement the basic tasks with the military flight simulator. Their age was form 22 to 31 and averaged 26. All of them had normal visual acuity with binocular acuity of 1.2 or better.

2.2 Apparatus

The study was conducted at a cockpit-based simulator with a high fidelity. Prototype of the simulator was a military fighter cockpit that replicated actual aircraft performance, navigation and dynamic system. The interior cockpit has a set of instrument panel, The real stick, rudder pedals and a throttle lever. Outside visuals were provided by a forward projection screen at a total field of view with 90 degrees.

Fig. 1. Helmet of Eyelink II System

Eye movement measurement was collected with an Eyelink II head-mounted eye tracking system which pilots could move their heads freely in flight simulation(Fig 1). The eye tracking system utilizes both pupil and corneal reflection and the data was sampled at 250Hz. The system's average gaze position error was less than 0.5°. There is a scene camera on helmet of the Eyelink II system that can record the video of the scene. Playback of the video can know the pilots' line of sight to any instrument and the sequence of instrument scan in the cockpit.

2.3 Flight Tasks

Participants were asked to fly three scenarios in visual flight rules(VFR) condition in a sunny day without wind with simulator. Scenario 1 involved landing phases and lasted approximately 50 seconds. Scenario 2 involved a level fly that lasted approximately 1 minute. Scenario 3 was a climbing phase and lasted approximately 40 seconds. The beginning place of three scenarios was the same. It was 400 meters high and 5400 meters away from the centre of the runway (Fig 2).

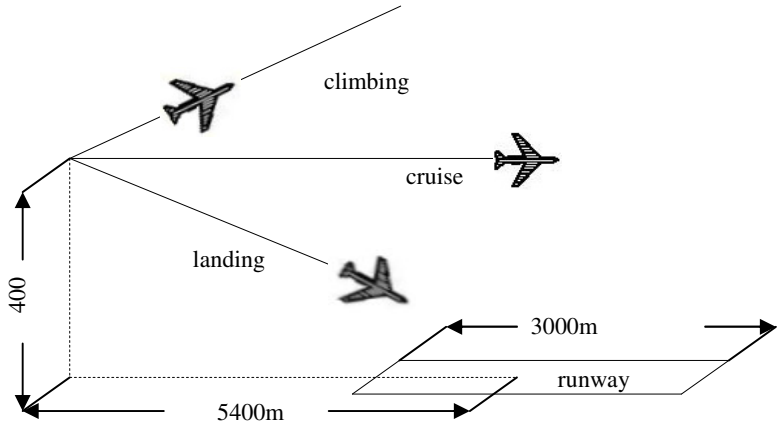


Fig. 2. Flight simulation task sketch map

2.4 Procedure

The actual experiment scene was shown as figure 3. At the beginning of each session, every participant received a brief introduction about the background of the study and experiment tasks. Before sitting into the cockpit, the helmet of the eye tracking system was put on the participant's head and calibrated. Each participant was allowed a five-minute familiarization to the flight and simulator. The simulator was then initialized to the position for flying.

Fig. 3. Actual experiment scene

3 Results

When pilots perform a flight mission, they should have different eye scanning mode and attention allocation strategy on out view(OV) and instrument panel(IP) of the cockpit. To study this difference, the vision information source was divided into two parts: area of interest(AOI) of out view, and area of interest of instrumentation panel of the cockpit. Five eye movement indexes which they were the percentage fixation point(PFP), the percentage dwell time(PDT), the average fixation duration(AFD), the average pupil size(APS) and the average saccade amplitude(ASA) were finally recorded and analyzed.

3.1 The Percentage Fixation Point

PFP is the ratio of fixation point number in each AOI to the total fixation point in each trail. The index was related to the fixation frequency to OV or IP. The result of PFP of three fly scenarios could been seen in table 1. The t Test showed that PFP spent on IP of cruise and climbing task had obvious difference($P<0.05$) to that of landing task. on instrument PFP. The difference between cruise and climbing phase was not significant ($P>0.05$).

Table 1. PFP in each AOIS of three flying stages(%)

AOI	Landing	Cruise	Climbing
OV AOI	82	69	66
IP AOI	18	31	34

3.2 The Percentage Dwell Time

PDT was a measure of the percentage of time that subjects looked at each AOI. PDT can quantitatively measure vision attention allocation to each information sources. PDT data (Table 2) showed that there exit great difference between OV AOI and IP

AOI. Compared to cruise and climbing phase, the time that subjects looked at OV AOI was more($P<0.05$) in landing phase. The time difference between climbing and cruise phase was not obvious ($P>0.05$).

Table 2. PDT in two AOIS of three flying stages(%)

AOI	Landing	Cruise	Climbing
OV AOI	97	84	80
IP AOI	3	16	20

3.3 The Average Fixation Duration

AFD was a ratio of dwell time in each AOI to the fixation point number in this research. Data (Table 3) shows that AFD in OV AOI was far more than that in IP AOI and was almost twice as much as that in IP AOI. AFD difference in OV AOI also exited between landing and cruise phase ($P<0.05$) and between climbing and cruise task ($P<0.05$) while the difference between landing and climbing task was not significant ($P>0.05$).

Table 3. AFD in two AOIS of three flying stages(ms)

AOI	Landing	Cruise	Climbing
OV AOI	594	501	586
IP AOI	250	220	276

3.4 The Average Pupil Size

APS data of three flying tasks was in table 4. T test results showed that APZ between OV AOI and IP AOI was of significant difference ($P < 0.05$).

Table 4. APS in two AOIS of three flying stages

AOI	Landing	Cruise	Climbing
OV AOI	1031	991	958
IP AOI	1085	1063	1012

3.5 The Average Saccade Amplitude

ASA of three scenarios was shown as figure 4. T test showed that ASA of landing stage and climbing stage were of significant difference($P < 0.05$) to cruise stage while there was no difference ($P > 0.05$) between landing stage and climbing stage.

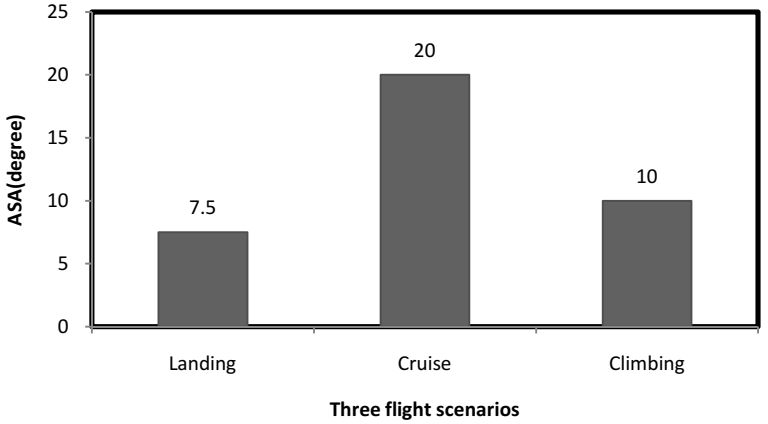


Fig. 4. ASA of three flight scenarios

4 Discussion

4.1 The Percentage Fixation Point

Two characteristics could be seen obviously from Table 1: the first was that the times which subjects fixated on out view was much more than they fixated on instrument panel; the second was that the fixation times to the instrument in cruise and climbing phase was more than that in landing phase. It also could be seen from the video playback that only the air speed indicator was viewed during the landing stage. The vertical speed indicator and altitude instrument were viewed more frequently during level flight stage while the vertical speed indicator, air speed and direction instrument were viewed more frequently during climbing stage. Some researchers found that the pilot viewed the altimeter more frequently during stages when heading was changing, And while heading and altitude were changing, the airspeed indicator was visited more frequently[4]. Their results were in the instrument flying rule (IFR) condition. The difference of fixation times indicated the change of scanning pattern on tasks and importance degree of the information sources to pilot's tasks. The optimal instrument layout or design can be reached by this important degree. For example, according to the times and direction of pilot's scanning, air speed indicator could be placed to the left side of the attitude indicator and altitude indicator is to the right while other indicator can be laid around them[1].

4.2 The Percentage Dwell Time

Data (Talble2) showed that there exit great difference between OV AOI and IP AOI. Subjects spent most of their time looking outside, which meant that subjects got information mainly from the out view, their attention was mostly allocated to out view. In realistic flights, aircraft controlling information mainly comes from speed, altitude, direction and attitude instrument. In VFR condition, pilots can acquire the most flying

information from the runway, the skyline, buildings and terrains, so they need not to pay more attention to instrument. Especially in the landing task, outside scene of the cockpit is so clear and change so quickly that the pilots are not able to look at instrument for much time. The data (Table2) showed a ratio of above 80% of the time allocated to the outside. In comparison, Wickens and colleagues found that pilots spent about 40% of their time attending to the outside world in cruise[5]. But it was in a general commercial aviation task and commercial pilots generally have more emphasis on instrument interpretation and rely much less on outside visual references than their military counterparts[6]. Similar to the data (Table 2), Peter Kasarskis found that roughly 13% of the time was spent on the instrument panel (87% outside) during landing phase[4], but the scenario is some different in the sort.

4.3 The Average Fixation Duration

The longer time of AFD attributed to two factors: the information was difficult to extract that required greater cognitive efforts, or that information was rich that need to spend greater time to read the various sources. Because there are skyline, runway and much more other visual reference information which pilot's fly needs, the longer dwell time might be a combination of both. The result of the longer time of AFD was the increasing cognitive workload. Landing and climbing task is more complex than cruise and workload during two phases is highly, so it was reflected on AFD. Especially during landing phase, pilots must get enough information and make accurate and complex information process for the precise landing. So this AFD difference may be the represent of the difficulty and complexity of the task. So it can be concluded that AFD is an index of task difficulty and workload and it will be longer with the increasing of the task difficulty and increasing of workload. This index also can be used to diagnose the readability of the cockpit display information; the well-designed display may need a shorter AFD.

4.4 The Average Pupil Size

Researchers believed that the pupil size was a sensitive indicators of workload. When participants tried to look at a target, the pupil size increased. In this experiment, there were two possible reasons with the increasing of pupil size: first, the position of the cockpit instrument was under the horizontal line of sight of subjects, they must make a downward looking to read the instrument; the second, Dark cabin environment made subjects increase the pupil size to obtain the instrument information. In addition, it was also a reason that subjects need to make continually fast instrument inspection and monitoring workload was improved. So it was suggested that the instrument should be tried to put in the optimum field of vision when make a layout design. The display form of instrument and ambient lighting should also make a reasonable design.

4.5 The Average Saccade Amplitude

It could be found that ASA had the relation to task difficulty. It agreed with the related study that ASA decreased with the increase of task difficulty[7]. A certain relationship existed between vision field and task difficulty. When task difficulty increased, the information content increased, then the vision angle of vision cone decreased, effective attention field of vision decreased, and the phenomena was named as "Tunnel Vision", so the ASA decreased. In landing phase, the pilot's visual scanning range was mostly on the narrow strip runway. Especially before the final earthing short moment, the pilot's vision was highly concentrated in a short runway of the front end of the airplane. The reason of Pilots' feeling of the task difficulty was that attention mechanism limited its access to more information, so as to protect the visual system overload.

4.6 Attention and Eye Movement

Other researchers have studied pilots scanning behavior with recollection report[8], however, since it reflected only what the observer remembered and most people were not always aware of where their eyes are looking at any given instant time. The measurement of eye movement provides a detailed history behavior of scanning patterns; eye movement is likely to be a more sensitive measure of the observers' intentional state. The fixation durations reveal sensitivity to stimulus changes that could not be reported verbally. Subjects can easily distribute their attention covertly across the visual field in the absence of eye movements. However, many studies supported the idea that the shifts in attention made by the observer were usually reflected by fixations[9]. In aviation domain, the eye movement behavior is task-driven and eye movement is mostly controlled by top down mechanism, so the fixation locus and focus of attention are tightly linked. The tight link between fixations and task performance in landing task also lends credence to the idea that fixations reflect the primary distribution of attention[4].

From above analysis, it can be thought of that eye movement indexes can measure pilots' attention quantitatively; different task difficulty and different dimension workload can also be indicated by eye movement measure. The index of the average fixation time can probe the state of cognitive workload. The pupil size is a sensitive index of visual monitor workload. The average saccade amplitude will change with the task difficulty. The increasing task difficulty improve the workload, the further increasing workload may lead to "vision tunnel" and reduce the attention extent. So the relation of task difficulty, attention and workload can be revealed by the measurement of eye movement.

5 Conclusion

Eye movement technology is an accurate and objective means of providing a window into a pilot's cognitive process. Through the analysis of this study, some conclusions can be made as follows: (1) Eye movement indexes are good indicators

of quantitatively measuring pilot's attention, the pilot's attention keeping, switching and allocation can be known by integrating multiple eye movement indexes. (2) Eye movement measures can reflect the different dimensions of workload and can make good diagnosticity to multiple dimensions workload. (3) The pilots' scanning mode to the out view and cockpit instrument is different. (4) In VFR condition, pilots' attention is mainly focused on out view. The information that they make decision to fly an airplane is mainly obtained from out view while they only have a quick crosscheck to the cockpit instrument.

Acknowledgement. This work is supported by the Technology Foundation of National Science (A0920132003), the Natural Science Foundation of China (31170895) and the opening foundation of the Science and Technology on Human Factors Engineering Laboratory, Chinese Astronaut Research and Training Center(HF2013-K-06).

References

1. Fitts, P.M., Jones, R.E.: Eye fixation of aircraft pilots, III. Frequency, duration, and sequence fixations when flying air force ground-controlled approach system. AF-5967 (1949)
2. Merchant, S.: Eye movement research in aviation and and commercially available eye trackers today (2001), http://www.lucs.lu.se/EyeTracking/overview.html
3. Thomas, A.F.: The super cockpit and human factors challenges (2003), http://www.hitl.washington.edu/publications/m-86-1/
4. Kasarskis, P., Stehwien, J.: Comparison of expert and novice scan behaviors during VFR flight. In: The 11th International Symposium on Aviation Psychology Columbus, The Ohio State University (2001)
5. Wickens, C.D., Xu, X.: The Allocation of visual attention for aircraft traffic monitoring and avoidance: baseline measures and implications for free flight. ARL-00-2/FAA-00-2 (2000)
6. Schnell, T.: Applying eye tracking as an alternative approach for activation of controls and functions in aircraft (2001), http://www.ccad.uiowa.edu/
7. Wu, D., Shu, H.: The application of eye movement measurement in the study of reading. Journal of Developments In Psychology 9(4), 319–324 (2001) (in Chinese)
8. Liu, W., Yuan, X.G., Liu, Z.Q., et al.: Experimental study of pilots' scan, and performance workloads. Space Medicine & Medical Engineering 18(4), 293–296 (2005)
9. Shinoda, H., Hayhoe, M.M.: What controls attention in natural environments? Vision Research 41 (2001)

Evaluation Research of Joystick in Flight Deck Based on Accuracy and Muscle Fatigue

Zheng Yang, Zhihan Li, Lei Song, Qi Wu, and Shan Fu*

School of Aeronautics and Astronautics，Shanghai Jiao Tong University, Shanghai 200240, China
sfu@sjtu.edu.cn

Abstract. Human factors have been the main reason of flight accidents, in which misoperation of pilots plays an important role. According to Statistics, in the field of accidents caused by human error accounted for about 80%. The layout design of aircraft cockpit and different controller locations result in different situations of muscle fatigue. Since the situation of muscle fatigue has an effect on the response accuracy, the probability of misoperations increases. The situation of muscle fatigue can be reflected by variables of the sEMG signal. This paper aims to verify the relationship between muscle fatigue and response accuracy based on analysis of sEMG signal. In the experiment, we investigated changes of response accuracy using joystick controller to trace the static object and dynamic object, along with the change of the situations of muscle fatigue. The terminal experiment result can provide information and design method of flight deck to make flight deck safer and more comfortable.

Keywords: human factors, muscle fatigue, response accuracy, sEMG signal.

1 Introduction

It is no doubt that safety is one of the most important criterions to evaluate Civil Aircrafts. From the first flight in December 17th in 1903, the accident rate due to the aircraft equipment failures has declined from 80% to 3% [1], as the technique of aircraft design and manufacture developed greatly. However, according to statistics, the error of flight crew reaches 1to 10 per hour [2], and in the field of aviation system error caused by human error accounted for about 70% to 90% [3-5]. Thus, human factors have been the main reason of fight accidents. However, the traditional ways to increase the reliability and safety of the aircraft can not solve these problems effectively. The research on aviation human factors has been brought to the forefront. [6]

Human-Centered Design can tackle the safety problem from its sources. Cockpits are the main places in which the pilots work, that's to say, majority of human-machine interactions occur in cockpits. Therefore, analysis of cockpit human factors in cockpit design is sure to be the key in aircraft design.

* Corresponding author.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 428–436, 2014.

The situation of muscle fatigue is closely linked with the cockpit comfort. [7] However, the recent design, lack of related muscle fatigue theory, can only refer to empirical data and general rules based on Ergonomics. Muscle fatigue of pilot is the important influencing factor of flight fatigue and cockpit comfort.

Joystick has been applied as controller in A320 which is the first application in civil aviation. Joystick can make the pilots control the aircraft more conveniently and quickly. Nowadays, joystick, as one of the major controllers, is a significant part of cockpit. The aim of this paper is to verify the relationship between muscle fatigue and response accuracy of controlling side stick based on analysis of surface electromyography signal (sEMG signal).

sEMG signal is biological electrical signal of neuromuscular system on the surface of muscles. There is some correlation between sEMG and the active status and function conditions of muscles. Thus, sEMG signal can reflect the situation of neuromuscular activity. The traditional method to analyze sEMG signal is to regard the signal as time function, use a few parameters, such as integral electrical values (Integrated EMG, IEMG) or statistical parameters such as Root Mean Square (Root Mean Square, RMS) to estimate muscle fatigue condition. According to research, when muscles begin to fatigue, the value of IEMG and RMS will increase. [8,9] And in spectral analysis, the method most often used for estimating the spectrum of the sEMG signal was the Fourier transform, partially due to the computationally effective fast Fourier transform (FFT) algorithm. [10] Some spectral parameters, such as Mean Power Frequency (MPF) and Median Frequency (MF) are normally used in analysis of sEMG signal. During static contractions, it has been shown that there is a strong correlation between spectrum characteristic parameters and muscle fatigue, the spectrum shift to the left as the fatigue processes, and the parameters MF and MPF go into a downward trend. [11-12] However, in dynamic fatiguing tasks, it is difficult to get a unified conclusion, because of large changes of MF and MPF. [13]

In this paper, we mainly use the time domain parameters IEMG and RMS, along with spectral parameters MF and MPF to estimate the muscle fatigue status. This paper focuses on the relationship between muscle fatigue and the response accuracy of controlling side stick.

2 Experiment Design

2.1 Participants

30 right-hand dominant volunteers (mean age 21±1.1 years), participated in this experiment. All participants are healthy, and had normal or corrected-to-normal vision. Before experiment, all subjects got enough sleep, and did not have strenuous exercises.

Before placing the measurement electrodes, placement site was identified. The electrode sitewas initially cleaned with sterile alcohol pads to by exerting asufficient abrasive action to avoid impedance mismatch and therefore improve the SNR. Motor points were locatedby means of a stimulator and the electrodes were positionedon the middle portion of muscle belly (short head) parallelto the longitudinal axis of muscle fibers and away from themain motor point. Apparatus

In this experiment, there are three main apparatuses. The first one is a computer, which was used to display visual stimuli on the screen and record the response accuracy. The 22-inch flat computer screen (1680×1050 pixels; 60Hz refresh) was fixed 70cm in front of the participants. An experimental program ran in the computer, to display two kinds of visual stimuli. The first kind is static stimulus, five straight lines, which were going to be traced, were showed on the screen all at once, these lines showing on the screen would last 4 seconds. The second kind of stimulus is dynamic stimulus, the line would appear on the screen every 2 seconds with random directions. There would be five straight lines in one group stimuli.

The second apparatus is sEMG signal measuring equipment—Trigno Hybrid Sensors (fully wireless,Trigno LAB, America 2000Hz), to measure and record the surface EMG signals form right biceps brachii, triceps brachii, and brachioradialis.

A joystick was fixed on a table and held by a participant's right hand, used to control the cursor on the screen. When the joystick was turned left or right, cursor on the screen moved along X axis, when it was pulled or pushed, cursor moved along Y axis.

All software implementations were done in MATLAB 7 with the Signal Processing toolbox 6.0, Statistics toolbox4.0, and Wavelet toolbox 2.2.

2.2 Experimental Procedure

Before the formal experiments, each participant was given a period of time to do the pre-experiments. In pre-experiments, participants were going to understand the experiment content, and be familiar with experimental process and operations.

The formal experiment was divided into two parts: static stimulus experiment and dynamic stimulus experiment. In the static stimulus experiments, volunteers were asked to trace 5 static straight lines on the screen. During the dynamic stimulus experiments, volunteers were asked to do three groups of dynamic stimuli trace, each group contained 5 straight lines. Between the two parts, participants were given five minutes to have a rest to make the fatigue has disappeared.

When the participants were tracing the stimuli on the screen, the computer would record coordinates of each line, and coordinates of trace paths. The surface EMG measuring equipment would record surface EMG signals during the experiments.

2.3 Data Processing

Analysis of Response Accuracy. In the experiments, we can get the coordinates of starting points and end points of the straight lines, and coordinates of trace paths, thus, we can calculate the average distance between the lines and trace path.

From the starting point (a, b) and the end point (c, d), we can get the equation of the line:

$$\frac{y-b}{x-a}=\frac{d-b}{c-a} \tag{1}$$

After simplified, get:

$$Ax + By + C = 0 \tag{2}$$

If a random point on the trace path is (x_0 , y_0), the distance from the point to the line is:

$$d = \frac{\left|Ax_0 + By_0 + C\right|}{\sqrt{A^2 + B^2}} \tag{3}$$

We calculated the average distance $\overline{d}$ of each point on the trace path. The value of $\overline{d}$ decreases, which means the response accuracy of controlling the joystick increases, vice versa.

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{4}$$

Analysis of Muscle Fatigue Status. In this paper, we mainly adopt time-domain analysis, along with spectral analysis to estimate muscle fatigue status.

Time-domain analysis is to regard the signal as time function, and then get statistical parameters through analysis, such as shaping and filtering the surface EMG signal, calculating the average rectified value and root mean squared of the signal, and use amplitude histogram, zero numbers, mean square value, the third order moments or fourth order moments as the signal characteristics for pattern classification. Time domain analysis is to describe the time sequence of the amplitude of the signal characteristics, mainly including integral electromyography (IEMG) and root mean square (RMS) value.

Integral electromyography (IEMG) refers to the sum of the area under the electromyographic signal obtained by rectifying filtering per unit time, which is an important means of evaluation of fatigue. Root mean square (RMS) is used to describe a period of time variation characteristics of average myoelectricity, which refers to all the RMS amplitude of this period of time, but it can't reflect the details of the electromyographic signal. Higher amplitude of fatigue electromyographic signal, is bound to cause an increase in the RMS. The time and the degree of fatigue can be identified by comparing RMS of different periods. According to many research results, from the initial state to the fatigue state, the general trend of the time domain value of sEMG is rising in the process of fatigue, which reflects the number of the working motor unit.

$$IEMG = \frac{\sum_{i=1}^{n}\left|emg_i\right|}{n} \tag{5}$$

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} emg_i^2}{n}} \tag{6}$$

Traditional spectral analysis method is to transform the time domain signal into frequency domain signal by Fourier Transform; Fast Fourier Transform (FFT) is commonly used in the signal spectrum analysis or power spectrum analysis. The power spectrum analysis of sEMG signal is widely used in muscle disease diagnosis and detection of muscle fatigue; the commonly used indexes are median frequency (MF) and mean power frequency (MPF). Most study in static fatiguing tasks shows that the sEMG power spectrum shift to low frequency when the muscles begin to fatigue, low frequency ratio increased, the proportion of high frequency decreases, MPF and MF would decrease. However, in dynamic fatiguing tasks, it is difficult to get a unified conclusion, because of large changes of MF and MPF.

$$MPF = \frac{\int_{f_1}^{f_2} f \cdot PS(f) \cdot df}{\int_{f_1}^{f_2} PS(f) \cdot df} \tag{7}$$

$$\int_{f1}^{MF} PS(f) \cdot df = \int_{MF}^{f2} PS(f) \cdot df \tag{8}$$

Thus, in this paper, we mainly use the time domain parameters IEMG and RMS, along with spectral parameters MF and MPF to estimate the muscle fatigue status.

3 Results

3.1 Static Stimulus Experiment

In the static stimulus experiments, we used MATLAB 7.0 to process the surface EMG signal, and got the tendency of four parameters (IEMG, RMS, MPF and MF) during experiments.

From figures above, we can find that IEMG and RMS, the two time-domain parameters had a trend of increase, which reflects muscles were in the fatigue state. However, two spectral parameters, MF and MPF did not have an obvious trend.

We can calculate the average distance between the trace path and target lines.

According to this figure, the average distance between the trace paths and target lines did not change a lot or decreased, which reflects that the response accuracy did not have an obvious change.

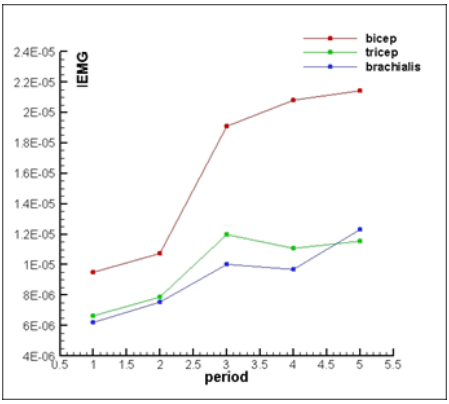


Fig. 1. IEMG in Static Stimulus Experiment

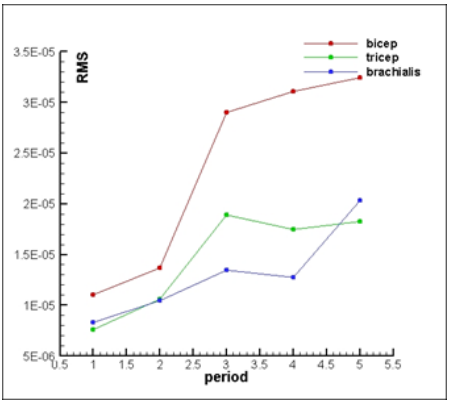


Fig. 2. RMS in Static Stimulus Experiment

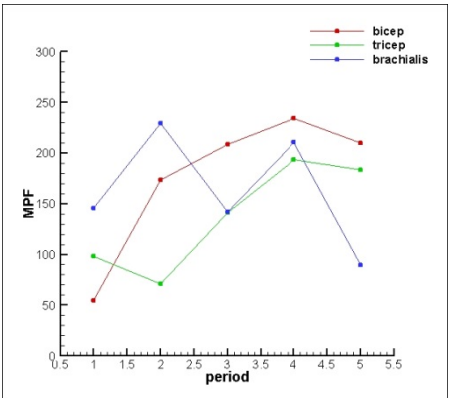


Fig. 3. MPF in Static Stimulus Experiment

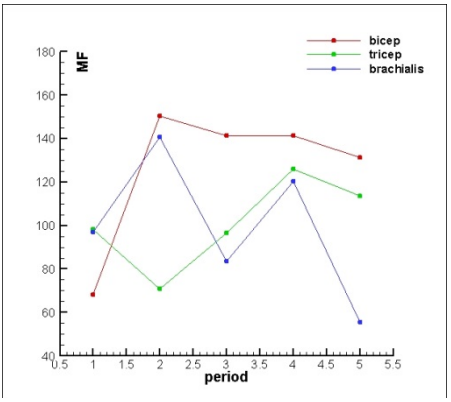


Fig. 4. MF in Static Stimulus Experiment

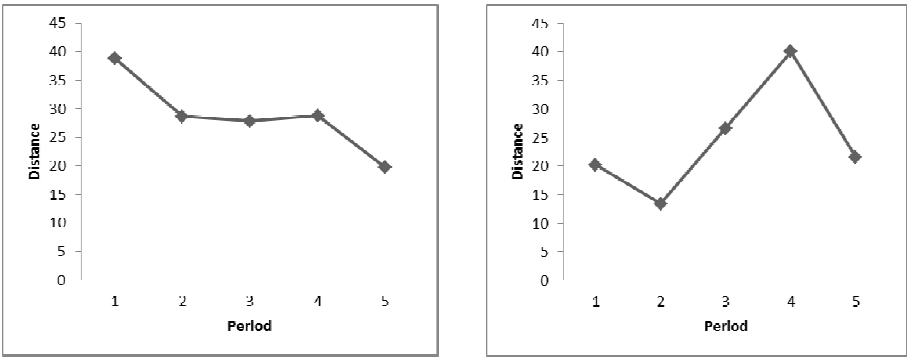


Fig. 5. Average Distance in Static Stimulus Experiment

3.2 Dynamic Stimulus Experiment

The same as the static stimulus experiment, we got the tendency of muscle fatigue parameters as follows.

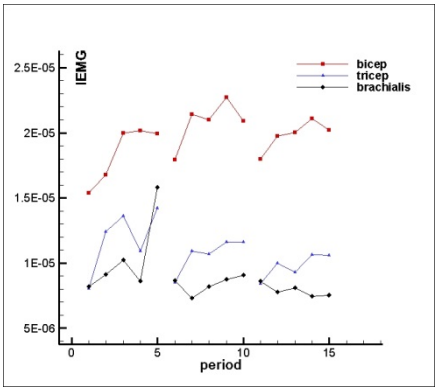


Fig. 6. IEMG in Dynamic Stimulus Experiment

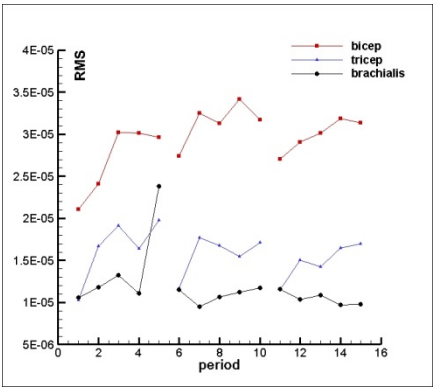


Fig. 7. RMS in Dynamic Stimulus Experiment

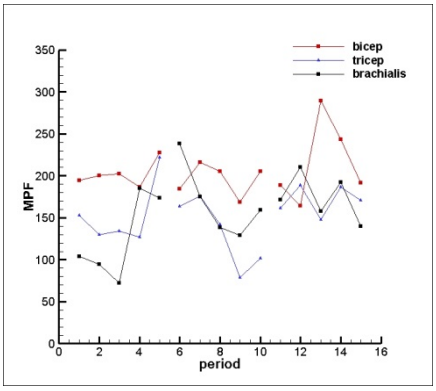


Fig. 8. MPF in Dynamic Stimulus Experiment

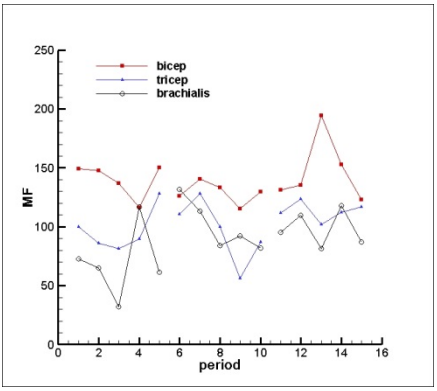


Fig. 9. MF in Dynamic Stimulus Experiment

According to these figures, we can find that IEMG and RMS, the two time-domain parameters had a trend of increase, which reflects muscles were in the fatigue state. However, two spectral parameters, MF and MPF did not have an obvious trend.

Then we calculate the average distance of each group, and got the following figure. In this figure, the average distance between trace paths and target lines increases. It reflects that the response accuracy declined.

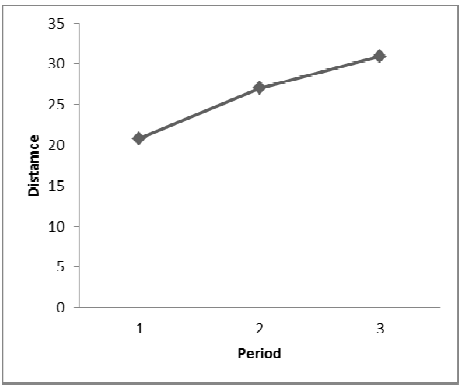


Fig. 10. Average Distance in Static Stimulus Experiment

4 Discussion

Considering the results in Section 3, we can identify that the time-domain parameters, IEMG and RMS, appeared the trend of increasing obviously. This trend reflects that muscles transformed from the normal state to fatigue state. However, the spectral parameters, MPF and MF, did not have an evident tendency. The level of sport intensity is too low (generally smaller than 20% MVC), the spectral parameters (MF and MPF) would not have apparent change.

In the static stimulus experiments, muscles began to fatigue, the response accuracy did not get a unified trend of increasing or decreasing. This situation verified that the muscle fatigue has little influence on response accuracy when people are faced with static stimuli. Some of the participants may have increasing response accuracy, due to they became more familiar with the operations as the experiment went on.

In the dynamic stimulus experiments, when the muscles began to fatigue, the response accuracy would decrease. The experiments verified the correlation between the muscle fatigue and response accuracy.

Because participants were asked to do the static stimuli experiments first, and then take the dynamic stimulus experiments, the operation proficiency may have an effect on the results.

5 Conclusion

In spite of the effect of operation proficiency, there is a correlation of the muscle fatigue and response accuracy. Especially faced with dynamic stimuli, human response accuracy of controlling side stick would decline, as the muscle began to fatigue.

The result of experiments verified that muscle fatigue has an influence on human response. People should pay more attention to the muscle fatigue in flight tasks, to prevent problems in aviation safety because of the low response accuracy.

Acknowledgements. This research work was supported by National Basic Research Program of China-(973 Program No. 2010CB734103).

References

1. The Civil Aviation Administration of China Human Factors Group. Civil Aviation Human Factors Training Manual. China Civil Aviation (March 2003)
2. Ding, X.: Research in the Effect of Flight Crew Human Factors on Flight Safety. Northwestern Polytechnical University (2005)
3. Ban, Y.: Aviation Accidents and Human Factors. China Civil Aviation (November 2001)
4. Summary of Commercial Jet A lane Accidents Worldwide Operations from 1959-2002. Boeing (2002)
5. Fatal Events and Fatal Event Rates by Airline. Since (1970), `http://airsafe.com/airline.htm` (revised January 14, 2005)
6. Zhou, X.: Strengthen Research in Human Factors, Improve the Level of Safety Management. Journal of Civil Aviation University of China (July 2002)
7. Zhang, E., Bi, C., Wang, G.: Experimental Research on Biological Signal of Muscle Fatigue under Dynamic Driving Environment. Chinese Journal of Engineering Design 1(4), 246–252 (2010)
8. Merletti, R., Lo Conte, L.R., Orizio, C.: Indices of muscle fatigue. Journal of Electromyography and Kinesiology 1(1), 20–33 (1991)
9. Farina, D., Merletti, R.: Comparison of algorithms for estimation of EMG variables during voluntary isometric contractions. Journal of Electromyography and Kinesiology 10(5), 337–349 (2000)
10. Medved, V.: Measurement of human locomotion. CRC Press (2002)
11. Stulen, F.B., De Luca, C.J.: Frequency parameters of the myoelectric signal as a measure of muscle conduction velocity. IEEE Transactions on Biomedical Engineering (7), 515–523 (1981)
12. Mannion, A.F., Patricia, D.: Electromyographic median frequency changes during isometric contraction of the back extensors to fatigue. Spine 19, 1223–1229 (1994)
13. Masuda, K., Masuda, T., Sadoyama, T., Inaki, M., Katsuta, S.: Changes in surface EMG parameters during static and dynamic fatiguing contractions. Journal of Electromyography and Kinesiology 9, 39–46 (1999)

The Research of Implementing SC to Evaluate Complexity in Flight

Yiyuan Zheng, Dan Huang, and Shan Fu

School of Aeronautics and Astronautics
Shanghai Jiao Tong University, P.R. China
{leodeisler,huangdan}@sjtu.edu.cn,
sfu@sjtu.edu.cn

Abstract. In aviation, the Standard Operating procedures (SOPs) provides typically a list of action items that allowing the pilots to complete tasks in flight environment. Therefore, the complexity of SOPs should be appropriate to guarantee the flight safety. In this paper, step complexity (SC) from nuclear power plant is introduced to evaluate complexity in flight in nine tasks selected from SOPs. The verification measurement of SC is difference of heart rate (HR-D) of pilots. From experiment result, SC is correlative to HR-D. However, the correlation is not significant enough. Thus, to evaluate complexity in flight efficiently, the SC measure should be modified.

Keywords: SC, HR-D, complexity.

1 Introduction

Task is defined as activities or actions that need to be accomplished within a defined period of time or by a deadline. It is ubiquity in daily life. The performance of a certain task not only influences the context circumstances, but also closely relates to the safety of the whole system. Especially, the omission of procedural steps in task is a forms of human error with serious consequences in many complex work settings (Reason 2002, Hobbs and Williamson 2003). For instance, several fatal crashes resulted from inadvertently omitting of the crew to set the flaps prior to takeoff and the warning horn malfunctioned (Degani and Wiener 1993). Therefore, to accomplish a task ideally, certain procedures should be followed step by step. The Standard Operating Procedures (SOPs) are designed as a series of step operations for pilot to deal with normal or abnormal conditions of the aircraft. However, how to evaluate step operations in SOPs is still controversial.

Step Complexity (SC) was supposed to be implemented in nuclear industry (Park, Jung, and Ha 2001, Park and Jung 2007), and (Xu et al. 2009) studied the influence of SC and presentation style on step performance in aerospace field. In both industries, SC shows an acceptable capability of complexity. In our study, we implement SC in flight circumstances to evaluate complexity in SOPs.

In order to verify the effectiveness of SC in flight circumstances, workload measurement has been used. As with increasing of complexity, the workload of operator

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 437–443, 2014.

increases simultaneously. It has been suggested that increase HR could be related with an increased workload (Mulder 1989). Therefore, we select difference of HR as an indicator.

In this paper, firstly, the method of SC is briefly described in section 2. The implementation results of SC in flight conditions and verification HR-D are compared in section 3. In Section 4 discussion of the study is shown.

2 SC Measure

2.1 SC Overview

SC introduces entropy concept into complexity measure in nuclear power plants (NPPs). SC includes two type of complexities which are logic complexity and size complexity from two graphs that are action control graph (ACG) and information structure graph (ISG). Action control graph contains step logic complexity (SLC) and step size complexity (SSC), which represent logical sequence of the required actions and the amount of required actions respectively, and information structure graph includes step information complexity (SIC), which indicates the amount of information to be managed. Two kinds of order entropy were used to describe the complexity. The first-order entropy is used to evaluate the regularity of the program control logic, and the second-order entropy can evaluate the number of hierarchical levels of the graph (Davis and LeBlanc 1988). A simple example graph of two kinds of order entropy is shown as following in Fig. 1. Considering first-order entropy of the graph, Node 4, 5 and 3 have same In and Out numbers (1 In with 1 Out), and same as Node 6 and 8 (1 In without Out) as shown in Table 1. Therefore, the first-order entropy $G_1$of the graph is calculated as:

$$G_1 = -\sum_{i=1}^{6} p(A_i) log_2 p(A_i) = 2.156$$

On the other hand, second-order entropy is calculated through considering neighbor nodes of each nodes. If the neighbors are same, then the nodes are organized as one class, the classes for second-order entropy is in Table 2. Thus the second-order entropy G_2 of the following graph is:

$$G_2 = -\sum_{i=1}^{6} p(A_i) log_2 p(A_i) = 2.750$$

According to different contents of complexities, SSC and SIC are obtained from second-order entropy, and SLC is from first-order entropy. In sum, SC is calculated by a weighted Euclidean norm as shown:

$$\text{SC} = \sqrt{(\alpha * SIC)^2 + (\beta * SLC)^2 + (\gamma * SSC)^2} \quad (1)$$

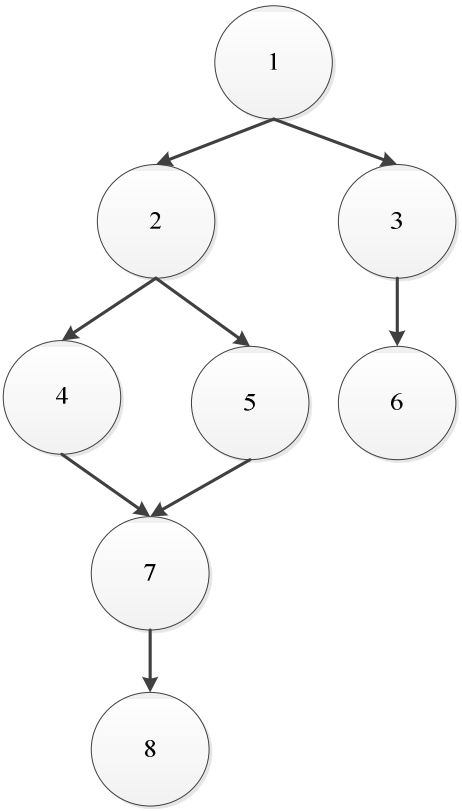


Fig. 1. An example graph for two kinds of order entropy

Table 1. Classes for first-order entropy

In	Out	Nodes	Class
0	1	1	1
1	2	2	2
1	1	4,5,3	3
2	1	7	4
1	0	6,8	5

Table 2. Classes for second-order entropy

Nodes	Neighbor Nodes	Class
1	2	1
2	1,4,5	2
3	1,6	3
4,5	2,7	4
6	3	5
7	4,5,8	6
8	7	7

2.2 Case Study

We selected final approaching phase from SOPs of Boeing 777 as an example. In final approach, in order to maintain descend as glide slope, pilot flying should establish approaching configuration of the aircraft above 1000 feet by controlling altitude, airspeed, heading and throttle. Meanwhile, he/she should confirm landing order from ATC. As operating procedures descriptions in SOPs, the first-order entropy equals to 2.807, and the second-order entropy is 2.522. In order to get SIC, more detailed analysis should be carried out on final approach. To settle correct configuration, the pilot should obtain information from prime flight display including altitude, airspeed and heading information. Moreover, he/she operates properly by manipulating control wheel, throttle and flight mode panel (FCP). At the same time, the pilot needs to communicate with ATC by setting frequency. Therefore, SIC is 3.322. From equation (1), SC equals to 2.903, where $\alpha = \beta = \gamma = 1/3$.

3 Validation Tests

3.1 SC Result

Nine tasks were chosen from SOPs including one engine failure, traffic collision avoidance system warning, the hydraulic system failure, etc. The results of SSC, SLC, SIC and SC of these nine tasks are shown in Table 3.

Table 3. The results of SSC, SLC, SIC and SC

Task	SSC	SLC	SIC	SC
Task 1	2.322	1.922	3.700	1.591
Task 2	2.000	2.000	3.000	1.374
Task 3	1.628	3.322	1.922	1.390
Task 4	4.437	1.928	4.858	2.285
Task 5	3.807	1.149	4.459	1.992
Task 6	2.322	1.371	1.685	1.060
Task 7	3.000	2.156	2.585	1.503
Task 8	3.700	3.085	3.459	1.978
Task 9	2.807	2.522	3.322	1.6758

3.2 HR-D Results

In order to obtain HR-D, same experiments as nine tasks were carried out in a Boeing 777 flight simulator as Fig. 2, and the participants of the experiments including 8 experienced pilots (mean=1965 flight hours, SD=932). Average HR-D were deduced

from the difference of HR value in tasks conditions and relaxation condition of pilots by a physiological parameters monitoring equipment (Bio Harness, Zephry Technology, Annapolis, U.S.A.). The results of HR-D is shown in Table 4.

Fig. 2. Boeing 777 Flight Simulator

Table 4. HR-D reults

Task	HR-D	Task	HR-D
1	5.15	6	4.16
2	8.23	7	9.22
3	3.53	8	12.18
4	18.71	9	23.42
5	15.12		

3.3 Comparison of SC and HR-D

As calculation and recording results of SC and HR-D, the correlation coefficient of these two value by Pearson correlation shown as Table 5. That means SC could somehow represent complexity in flight.

Table 5. The correlation coefficient of SC and HR-D

		HR-D
SC	correlation coefficient	.689*
	Significance	.040

Besides Pearson correlation, the exponential regression curve of SC with HR-D is displayed in Fig. 3.

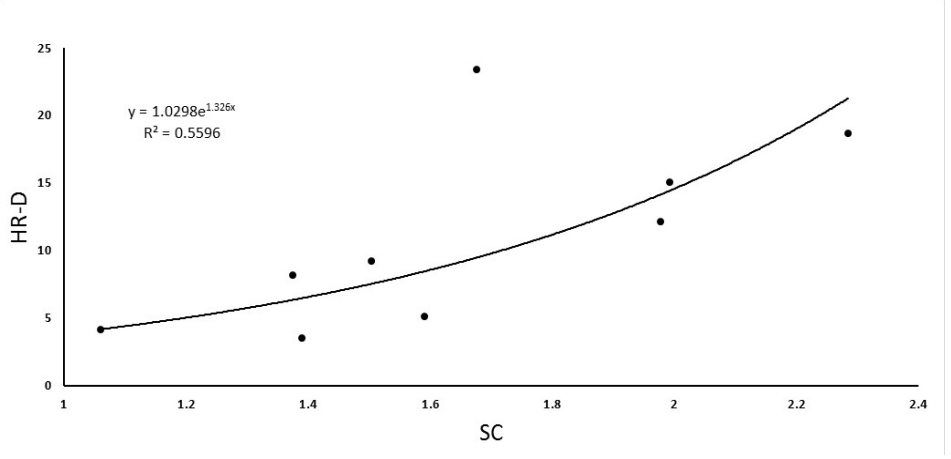


Fig. 3. The exponential regression of SC with HR-D

4 Discussion

In this paper, the preliminary study of quantitative indication of complexity in flight is carried out. This is important to aviation safety. Because excessive workload might result in disasters (Reason 2002). SC measurement from nuclear industry was introduced to represent the complexity.

From the experiment result, SC measurement is correlative to heart rate change. It could be considered related to complexity. However, the correlation is not significant enough. Although similar with nuclear plant, in flight much more operations are required to maintain the aircraft flying, not only surveillance tasks. Considering in real flight conditions, the real world operations might perturb and disrupt the executions (Greeno 1989), and these disruptions might cause the attention shift from the current task of the pilot. Therefore, more significant representation of complexity in flight is necessary.

According to the specific environment of flight, pilots might have much information exchange with the aircraft current configuration. Meanwhile, the operations are more complicated than in nuclear plant, different actions might yield the same results, or same device might have multiple functions. In further study, the above factors should be considered to form a more significant indication of complexity in flight.

5 Conclusion

In our study, we implemented SC from nuclear planet to evaluate nine flight tasks, and the verification method is difference of heart rate. The results of the experiments shows SC could indicate complexity in some extent. Nevertheless, the correlation might be improved by considering flight circumstance in further study.

Acknowledgements. This research work was supported by National Basic Research Program of China-(973 Program No. 2010CB734103).

References

1. Davis, J.S., LeBlanc, R.J.: A study of the applicability of complexity measures. IEEE Transactions on Software Engineering 14(9), 1366–1372 (1988)
2. Degani, A., Wiener, E.L.: Cockpit checklists: Concepts, design, and use. Human Factors: The Journal of the Human Factors and Ergonomics Society 35(2), 345–359 (1993)
3. Greeno, J.G.: Situations, mental models, and generative knowledge. Complex Information Processing, 285–318 (1989)
4. Hobbs, A., Williamson, A.: Associations between errors and contributing factors in aircraft maintenance. Human Factors: The Journal of the Human Factors and Ergonomics Society 45(2), 186–201 (2003)
5. Mulder, L.J.M.: Cardiovascular reactivity and mental workload. International Journal of Psychophysiology 7(2), 321–322 (1989)
6. Park, J., Jung, W.: A study on the development of a task complexity measure for emergency operating procedures of nuclear power plants. Reliability Engineering & System Safety 92(8), 1102–1116 (2007)
7. Park, J., Jung, W., Ha, J.: Development of the step complexity measure for emergency operating procedures using entropy concepts. Reliability Engineering & System Safety 71(2), 115–130 (2001)
8. Reason, J.: Combating omission errors through task analysis and good reminders. Quality and Safety in Health Care 11(1), 40–44 (2002)
9. Xu, S., Li, Z., Song, F., Luo, W., Zhao, Q., Salvendy, G.: Influence of step complexity and presentation style on step performance of computerized emergency operating procedures. Reliability Engineering & System Safety 94(2), 670–674 (2009)

Transport and Industrial Applications

Attending to Technology Adoption in Railway Control Rooms to Increase Functional Resilience

Elise G. Crawford[1,2], Yvonne Toft[1,2], and Ryan L. Kift[1,2]

[1] Central Queensland University, Rockhampton, Australia
[2] Centre for Railway Engineering, Rockhampton, Australia
{e.crawford,y.toft}@cqu.edu.au, rlkift@hotmail.com

Abstract. Introducing new train traffic management technologies can activate undesirable changes to operational safety. Therefore, it can be useful to understand how to expedite technology adoption in control rooms to strengthen the collaborative efforts of the human-automation team in support of resilient processes. This exploratory study presents factors that impact technology adoption. Results revealed that end-user buy-in was considered critical. Participants revealed that buy-in can be undermined when end-user expertise is not or under utilised and when horizontal communication channels are restricted. Technology issues arise when end-user work needs are not supported and when insufficient time, training or support slow adoption processes. Finally, organizational factors included: weak commitment and leadership to resource and drive project processes and dishonesty and lack of open accountability. Finally, stakeholders recognized that new projects are frequently managed from the top-down and that contributions from the bottom-up can add significant advantages toward expediting system changeovers.

Keywords: resilience engineering, technology adoption, control, sociotechnical, human factors.

1 Introduction

System complexity seems to be a natural progression in today's technological climate. However, although technology advancement can offer unique solutions their added complexity can also make it harder to manage operational safety, such as maintaining undisrupted train traffic. This is primarily because the railways belong to the class of sociotechnical systems where system functioning exists from the coordinated efforts of all components within that system [1]. Furthermore, unsafe events in the railways can lead in far reaching effects. A few days ago a commuter train killed 13 passengers as it smashed into a shuttle bus in Ukraine [2]. Months earlier, in North Dakota, a freight train carrying crude oil collided with another train sending fireballs 100 feet into the air causing the evacuation of 2,400 nearby residents [3]. For these reasons, organizations like the railways continually seek ways to improve their ability to bounce back from disturbances and maintain operational safety.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 447–457, 2014.

Increased system complexity and its impact on safety concerned many industries, particularly those organizations where safety is paramount. Concerns lead to an examination of organizations that had high safety success despite the presence of high operational disturbances. Examples in the study included aircraft carrier operations, commercial air traffic control systems, and nuclear power plants. These organizations were described as High Reliability organizations for the following reasons: they were sensitive to operations, reluctant to simplify explanations for accidents, preoccupied with failure, they deferred to expertise and were considered resilient [4].

The achievement of resilience has become the new way of thinking about how safety should be managed, particularly for organizations that experience constant pressure to increase productivity while maintaining extremely high-levels of safety. A resilient organizations is said to be one that can manage the unexpected and is also able and prepared to cope with surprises [5]. This approach has been labelled resilience engineering otherwise known as organizational resilience. The definition of a resilient system has developed over the last few years and is currently defined as:

"[having the ability to] adjust its functioning prior to, during, or following changes and disturbances, and thereby sustain required operations under both expected and unexpected conditions" [6].

Recent studies have found that for resilience to be effective, the organization must contain four essential qualities: (1) the ability to respond effectively, (2) the ability to monitor performance variability, (3) the ability to anticipate disruptions and pressures and finally, (4) the ability to learn from their experience [7]. The investigation for improved safety practices also revealed that system failures are rare primarily because people are very adaptable and through creativity and problem solving abilities operators are continually making adjustments to ensure system functionality. This constant performance variability has been found to be a normal phenomenon in sociotechnical systems and contribute to the high levels of successful functionality. However, the insufficient or inappropriate adjustments were found to be the performance variations that led to unsafe states [8]. Recognizable periods of increased performance variability have been found to exist when new technologies are being introduced into complex working environments. Unfortunately, history has shown us that many new technologies have contributed to unwanted events and prevented timely interventions [9].

In light of continual system performance variance and the need for strengthened resilience, it can be useful to understand how to expedite the technology adoption process when new traffic management systems are introduced into railway control rooms, a period known to evoke functional adjustments and performance variability. Therefore, the aim of this study was to determine ways to build organizational resilience during times of known functional disturbances, such as when new technology is being introduced. The study explored ways to expedite technology adoption in railway control rooms with the objective to strengthen the collaborative efforts of the human-automation team in support of resilient processes.

2 Methods

Two data collection methods were undertaken concurrently, a survey questionnaire containing two open-ended questions and a semi-structured 30 minute interview protocol. The purpose of the survey was to allow for analytical comparisons across control room technology stakeholder groups. Results that the indicate similar and conflicting opinions may help to enlighten which factors are a concern and for whom. These concerns might then be targeted and addressed to assist the relevant stakeholders. The two open ended questions required participants to comment on (1) the reasons why newly introduced technologies fail and (2) to identify factors that help new technologies to succeed. The purpose of the interviews was to allow control room operators an opportunity to directly comment on matters that impact their ability to maintain safe operations.

Random purposive sampling was achieved allowing the study to obtain comments from the desired population set, that being control room technology stakeholders. Three hundred and fifteen stakeholders completed the two open-ended questions. Stakeholders comprised of *Managers* (technology clients), *Designers* (technology builders and suppliers), *Evaluators* (human factors and safety professionals) and *End-users* (technology operators); while 37 control room operators from three safety-critical industries (network rail, air traffic control and power processing) were interviewed. Results from both data sets were triangulated to validate or dispel the acknowledged statistical results.

3 Results

3.1 Surveys

Of the 315 participants who responded to the two open-ended questions, stakeholder groups comprised of: 48 Managers, 85 Designers, 114 Evaluators and 68 End-users. Responses were sorted according to factors identified. Seven factors emerged from three broad areas: organizational culture, technology and the end-user (control room operator). Direct comparisons were conducted between stakeholder groups. To make data comparison possible between stakeholder groups that contained different numbers of participants, response figures per stakeholder group were converted to 'percentage of response' per stakeholder group. Figure 1 shows that all stakeholder groups identified issues in all seven areas. Results were fairly similar across stakeholder groups, with the greatest difference of opinion existing between managers and designers over the design and build processes.

Although the product outcome and implementation were areas of great concern across almost all stakeholders, achieving operator buy-in was identified as the greatest concern, by all stakeholder groups except for management. Managers expressed greatest concern over the design and build processes. Some interesting disparities between stakeholders show that, unlike managers, designers were less concerned with the design and build process as a potential failure point. Furthermore, aside from

identifying the achievement of operator buy-in as a primary obstacle, designers felt that the product outcome and its implementation were two areas of potential threat to success. Both designers and end-users had higher concerns regarding the achievement of a suitable concept idea, and all stakeholder groups identified potential issues regarding aspects of feasibility, particularly over: safety concerns and sufficient resourcing for post implementation activities.

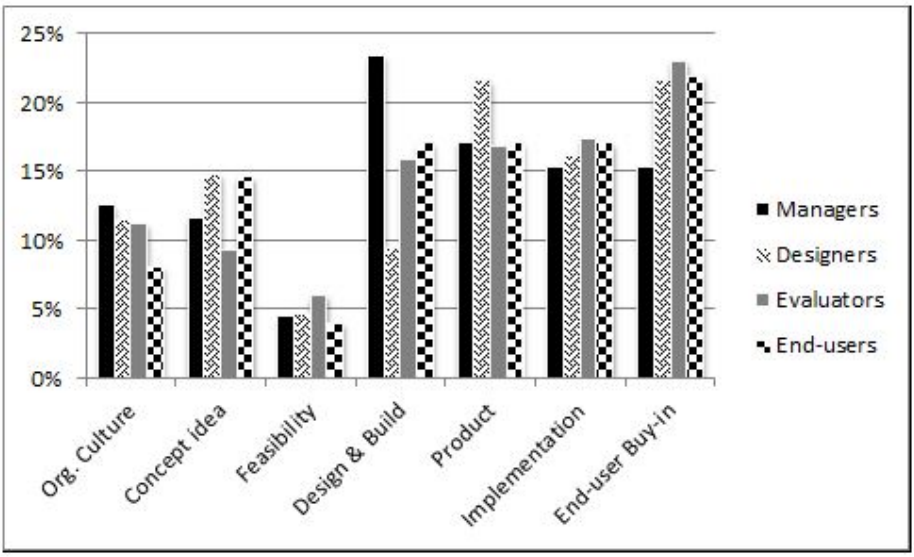


Fig. 1. Factors influencing technology adoption per stakeholder group according to percentage of response

3.2 Interviews

The factors for concern that emerged from the interviews closely aligned with factors revealed from the survey results. Figure 2 models the areas of concern within the three components of a sociotechnical system: with organizational culture playing an important role as decision maker, critical to adoption success falls with the end-user at the core, and integrated throughout the project development lies the technology design and implementation process.

Results from the interviews, revealed that an organizational culture of inconsistent values was responsible for many of the difficulties operators experienced. Poor decision making was identified to impact system functioning throughout its lifecycle, making each stage of the new technology project vulnerable to success for the following reasons: not achieving the right design concept from the outset, not ensuring that the new technology project was feasible in both financial and safety terms, deficient design and build processes that failed to involve end-users at pertinent stages, inadequacies in the final product outcome, and issues and problems often unresolved at implementation. Participants identified that core to achieving project

success was the achievement of end-user buy-in. Consequently, a perfectly good project, in the eyes of engineers and managers has the potential to fail if end-user buy-in was not achieved.

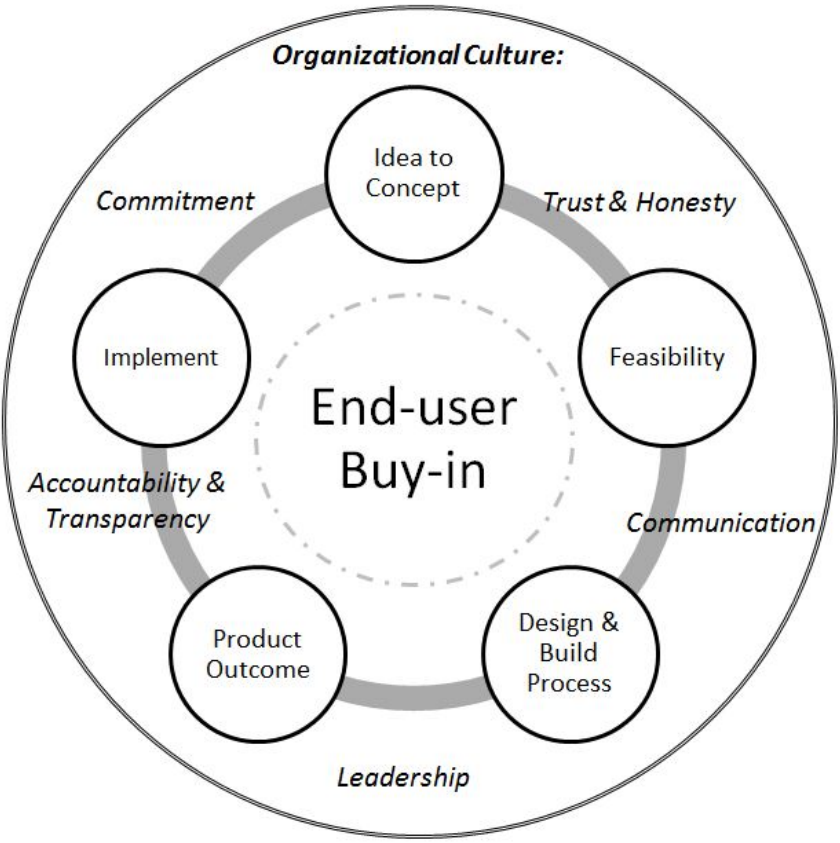


Fig. 2. Factors influencing successful technology adoption according to control room operators

Control room operators admitted that they, as end-users, ultimately determine the level of success of a newly introduced technology. Furthermore, they revealed that buy-in impacts their desire and likelihood to adopt the new system, stating that if they resist and do not adopt the technology, then the project is bound to fail. Participants also acknowledged that buy-in is only the symptom of other factors and therefore can only be improved if the other factors are addressed.

Interview participants suggested that organizational issues were found to affect the decisions made at each stage of the system design lifecycle. Major organizational factors identified to impact system design were primarily assigned as a responsibility of top management. These included: *trusting* your domain subject experts, the end-users; being *honest* when communicating information and conducting business; encouraging *communication* transfer both vertically and horizontally, thus allowing

controllers to discuss the pros and cons of the new system with trusted peers; being *committed* to the project and to system performance; exhibiting *leadership* by keeping the project on track and by mobilizing resources to ensure project success; and finally, to alleviate suspicion and uncertainty within the business by being *accountable* and *transparent* regarding actions and decisions made in all business affairs, particularly those that impact significant others.

4 Discussion

Comparisons between the two sets of results, found a close alignment of opinion regarding the factors identified to impact technology adoption during new technology projects. This supports the notion and advice frequently provided by safety experts who recommend addressing human factors at each stage of the system lifecycle to ensure safety [10-12]. Some of the major concerns identified are discussed below:

4.1 End-User Involvement to Get the Right Idea and the Right Design

Concept stage. Controllers expressed great concern over decisions made that impact their ability to maintain safe control. They expect that, if management is going to charge them with an area of responsibility, then they need to be provided with the tools to be able to meet their job demands. Unfortunately, interviews revealed that control room operators, including respected senior personnel, were rarely included in major projects. Operators want management to trust their expert contribution regarding their work domain. Control room operators and their technical maintenance staff are the subject experts for the tasks they complete within the control room environment and they would like to be recognized and respected as such and consequently be consulted in matters that impact them. Therefore, operators indicated that they were not interested in being involved in all aspects of the design process. Matters operators were interested in being included: idea to concept stage to ensure the idea was the right idea, one that meets their operational needs and system functional compatibility.

Consulted on feasibility. Operators want to be consulted before financial decisions are finalized to ensure the new technology can be implemented without undermining safety. They want to know that there will be sufficient resources allocated to help them become familiar and skilled with the new technology. They also want sufficient funding to ensure fine tuning the product activity can take place post implementation. One might not expect operators to be concerned about feasibility studies. However, experience has taught them that 'no new technology will be perfect'. Managers on the other hand, feel financial concerns are their responsibility and often view operator complaints unwarranted.

Algorithmic issues not their concern. Operators did not want to be involved in the internal, algorithmic technical build stage, which one designer suggested. Rather,

operators want to help designers to ensure that the interface provides the necessary information and feedback that is compatible with their needs to ensure optimal system functionality is achievable, particularly during contingency work.

Training. As the project draws close to commissioning, operators expressed a desire to have their design representatives (trusted peers) trained on the equipment and to be trained to train others. Operators noted that to be trained by work colleagues, allows for realistic real-world and site specific scenarios to be addressed, practiced and tested, making the learning experience far more applicable to their circumstances and thus more valuable. Although cost effective from a management point of view, operators expressed an aversion to on-line training for new control technologies. Rather, operators prefer to try the new technologies out, test them, and find out for themselves whether the system is robust or not. Through this familiarization process, controllers stated that they learn the strengths and weaknesses of the new technology, something online learning cannot facilitate.

Post implementation. Operators believe that a great deal of patience is required of them post implementation while they identify areas that can be adjusted or fixed to refine the new technology. As mentioned earlier, fine tuning new technology requires commitment from management to ensure recommendations are noted, negotiated, resourced and actioned to ensure their efforts make an impact. Many controllers expressed their emotional fatigue, stating that, after a while, if suggestions and improvements are not taken into consideration, or if they do not receive feedback and hence have no way of knowing whether their input has been considered, they simply lose interest. In time, they no longer bother to make the effort to report imperfections and lose the desire to offer further recommendations in preference to continued work-arounds and functional inefficiencies. These performance variances lead to impaired sensitivity and a lack of timely response and can adversely impact aspects of organizational resilience. The take away lesson here is, to ensure feedback is provided to end-users regarding all suggestions made, and preferably to devote some time for discussion to ensure that a certain level of appreciation and respect is afforded to operators as subject experts deserve.

4.2 Subject and Domain Expertise

Operators believe they can make significant contributions toward the improvement of operational safety. Human factors professionals support this notion. Operators explained that they are the only ones who are intimate with their systems and this makes them the best people to identify their needs toward achieving safe operations. They feel they can help to identify the appropriate provision of information and feedback that helps them manage system deviations. Furthermore, operators want to make improvements, they want to maintain safety, and they want to be effective at their jobs. However, as identified above, if suggestions are continually ignored, the operators become disillusioned and desensitized to potentially damaging deviations which eventually undermine and weaken organizational resilience.

Desire to be represented appropriately. Concerns existed around who was chosen to represent user needs and requirements. Control room operators consider it a lack of genuine commitment to the success of the project when management do not believe that end-users' can make a useful contribution. Having sufficient time away from the control room to take up opportunities offered to them was identified as a major obstacle. Operators expressed great frustration and disappointment when provisions were insufficient to ensure appropriate staff replacement or when meetings were scheduled at times that were in conflict with shift work designs. Operators noted that individuals who have not worked in the control room for some time, and supervisors or management who believe they know what the control room needs are, are all inappropriate representative choices. Unanimously, operators felt that it was not necessary to include all end-users in the consultation sessions, nor did all controllers express a desire to be involved. However and importantly, they all agreed that a need exists for at least one, but preferably more, trusted peers who can represent the needs within the control room faithfully. Therefore, it can be a mistake to assume that anyone other than those working within the immediate system has enough intimate knowledge to be able to represent significant others.

Furthermore, representation too late in the design process has support from recent investigations whereby new research from the human factors community is finding that end-user consultation at the very earliest possible stages is highly beneficial regarding cost, time, and value [13]. Furthermore, as interviewees pointed out, once specifications have been finalized at the concept stage and the product has gone to tender, there is very little a controller can do to ensure the new technology will improve his or her ability to maintain safe operations.

The opinion of many designers and managers was that end-users' do not really know what they want and that their inclusion in new projects can lead to the death of the project. Progress delays, budget blowouts and differences of opinion between operators are the primary reasons *why* end-users are frequently and deliberately not involved in new projects. Unfortunately, waiting until user acceptance testing to engage operators is regarded as too late for them to make any significant improvements which build frustration and animosity between controllers and their managers.

4.3 Elaboration on Factors that Impact Technology Adoption Success

A representation of factors identified at each stage of the system lifecycle is reproduced in Table 1.

Table 1. Factors at each stage of the system lifecycle that can impact system reliability

Stage	Issue that can impact system reliability
Idea to concept	Insufficient communication with all stakeholders (i.e., those impacted by the new idea)
	Insufficient consultation with subject experts including users and maintenance technicians
	Not addressing end-user priorities
	Not achieving a concept that meets end-user needs
	Not ensuring essential user needs are not compromised out
	Lack of communication both from management and across end-users
Feasibility	Inadequate resources to ensure project success (i.e., time, staff, training, fine tuning)
	Lack of attention to the new product impact (i.e., safety, compatibility, applicability)
Design & Build	Consultation with end-users too late to make a real difference
	Lack of end-user involvement during the testing stages
	Insufficient leadership from upper management to support user involvement
Product	Products that do not resolve a pressing need
	Products that do not offer a recognizable benefit
	Complex products that are difficult to use or are not user friendly
	Products that create more work or require major work-arounds
	Products that do not have problems resolved
	Products that to not perform as they were designed to perform
	Products that are incompatible with existing systems (human and technical)
	Products that restrict or impair maintenance work
Implementation	Lack of user familiarization and experience with the product
	Lack of hands-on training, and continued learning and technical support
	Lack of continued support from the manufacturer to refine the product

5 Conclusion

In conclusion, this study has shown that control room technology stakeholders recognize that system optimization and consequential technology adoption can be undermined in seven areas: at the top involving organizational issues, at each of the 5

major stages of the system design lifecycle, and with the operator in regards to end-user buy-in. Furthermore, operators acknowledged that they resist technologies that do not improve their ability to manage operational safety. Therefore, to improve technology adoption and utilization, operators offered a number of ways forward. Two ways identified to improve technology adoption and system efficacy included: (1) involve operators early in project development to help minimize the creation of unnecessary problems that may propagate during system development and implementation, and (2) trust operators as domain experts to provide the necessary feedback that can help system developers build products that support their situational awareness needs, maintain their mental models and support decision making.

Importantly, the processes required for effective action must start at the top of the organizations. In this way, management can make a significant impact by demonstrating commitment and leadership by establishing, resourcing and maintaining processes that appropriately utilize subject expertise. Such processes include: vertical and horizontal communication channels, consultation with and participation of subject experts (the operators) on matters that impact them, and by conducting business honorably and transparently to demonstrate accountability thereby reducing potential misunderstandings that undermine end-user buy-in.

Finally, although railway organizations have been in business for many years, this study has shown that new projects are frequently managed solely from the top-down, and that significant contributions made from the bottom-up are being underutilized. It is therefore recognized that the international railway community could benefit from the sharing of lessons learned from organizations with successful projects and/or the sharing of practices used by those with robust track records. The UK Network rail is one such organizations who has achieved successful project results of note and worth further examination [14]. These lessons can help to strengthen organizational resilience throughout the railway industry. Finally, today's technology market is a global concern and, therefore, international collaborative projects can have an expansive impact. As this paper has illustrated, end-user participation during new technology projects may increase organizational resilience and is therefore a worthwhile area to further investigation and validation.

Acknowledgements. The authors wish to thank all participants for their contributions and in particular the following organizations: Air Services Australia: Rockhampton, Brisbane and Melbourne, Stanwell Power Station; Metrol Rail Melbourne, Queensland Rail Rockhampton, and Ergon Energy Rockhampton.

References

1. Dekker, S.: Drift into failure. Ashgate Publishing Limited, Farnham (2011)
2. RT Network: 13 killed as commuter train rips apart shuttle bus on crossing in Ukraine (February 7, 2014), http://rt.com/news/train-crash-ukraine-b

3. Walsh, B.: North Dakota derailment shows dark side of America's oil boom (February 7, 2014), http://science.time.com/2013/12/31/north-dakota-rail-accident-and-oil-shipping-danger/
4. Weick, K.E., Sutcliffe, K.M.: Managing the unexpected, 2nd edn. Jossey-Bass, San Francisco (2007)
5. Woods, D.: Resilience engineering: redefining the culture of safety and risk management. Human Factors and Ergonomics Society Bulletin 49(12), 1–3 (2006)
6. Hollnagel, E.: Resilience engineering as an approach to safety for industry and society (2013), http://symbio-newsreport.jpn.org/files/upload/report/presentation_1365003000.pdf (February 4, 2014)
7. Hollnagel, E., Woods, D., Leveson, N. (eds.): Resilience engineering: concepts and precepts. Ashgate Publishing Limited, Aldershot (2006)
8. Hollnagel, E.: Resilience engineering in a nutshell. In: Hollnagel, E., Nemeth, C.P., Dekker, S. (eds.) Resilience Engineering Perspectives. Remaining sensitive to the possibility of failure, vol. 1, Ashgate Publishing Limited, Aldershot (2008)
9. Casey, S.: Set phasers on stun and other true tales of design, technology, and human error, 2nd edn. Aegean, Santa Barbara (1993)
10. International Organization for Standardization. Ergonomic design of control centres. Part 7: Principles for the evaluation of control centres. ISO, Geneva (2006)
11. Bahr, N.J.: System safety engineering and risk assessment. Taylor & Francis, New York (1997)
12. National Standards Authority of Ireland. Railway applications - The specification and demonstration of Reliability, Availability, Maintainability and Safety (RAMS) Part 1: Basic requirements and generic process. NSAI, Dublin (1999)
13. Rail Safety & Standards Board. Understanding human factors: a guide for the railway industry. RSSB, London (2008)
14. Network Rail. Britain relies on rail (February 6, 2014), http://www.networkrail.co.uk/aspx/662.aspx?cd=8

The Contribution of Automation to Resilience in Rail Traffic Control

Pedro NP Ferreira[1] and Nora Balfe[2]

[1] DREAMS-ULHT/ISLA Santarém, Portugal
ferreira.pnp@gmail.com
[2] Trinity College Dublin, Ireland
balfen@tcd.ie

Abstract. This paper addresses the challenges of high system complexity within rail traffic control. Based on resilience engineering principles, the different types of traffic control technology are analysed in order to identify either their contributions or hindering factors towards system resilience. Throughout four main generations of technology in traffic control, whilst there is a clear path towards increased automation, evidence from recent research in this domain suggests that the introduction of automation does not necessarily contribute to enhanced resilience. Despite its contributions to efficiency by placing larger areas under the supervision of each control post, it has introduced many new complexities in traffic control decision making. In many cases, automation has created a gap between rail operations and those in charge of their control. Beyond basing their decisions on operational needs and priorities, Traffic Controllers must take into account the possible responses that automated systems might initiate. So far, traffic control technologies are unable to deal with disruptions and much of the variability inherent to complex operations such as the railway but future generations of rail signalling systems may be able to better support resilience if appropriately designed.

Keywords: Complexity, decision making, flexibility versus rigidity, ETTO, automation, resilience, rail human factors.

1 Introduction

Technology is today profoundly embedded in every aspect of our social and economic activities. In a few decades, it has produced major transformations across all industry domains, both at management and operational levels. Within a safety management context dominated by human error concerns, automation has been a source of increased efficiency and quality (higher outputs and standardisation), whilst shifting human action away from the "sharp end" towards progressively higher and more complex levels of systems control and supervision.

More recently, in the domain of traffic control, the widespread application of information and communication technologies has tackled some of the challenges resulting from the highly distributed decision making processes, on which such control and

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 458–469, 2014.

supervision tasks are grounded. Within rail traffic control, there is evidence to suggest that, in addition to coping with increasingly dynamic and large scale operations, new system complexity related factors are emerging as a result of heightened automation, which may significantly hinder the ability of humans to cope with traffic control demands (Balfe et al., 2012). For example, the need for the operator to consider which factors have been accounted for by the automation and which have not. While discussing the impacts of complexity within most currently existing systems, Leveson (2004) points out that the increasing presence of software at all levels of management and operation gave way to "more integrated, multi-loop controls in systems with dynamically interacting components". This generates system interactions increasingly difficult to understand and control.

Based on the concept of resilience and in particular on resilience engineering as a framework for coping with high complexity in sociotechnical systems, this paper develops an overview of the evolution of automation in rail traffic control and investigates its impacts on decision making and overall system performance. After briefly introducing the key aspects of resilience engineering, a description of rail traffic control is developed, highlighting the main generations of systems, which over the years, have supported it. The impacts of automation on decision making processes are then discussed in light of resilience engineering literature, with particular interest on aspects which may hinder or enhance system resilience.

2 Resilience

Resilience has become a widely used concept across many different domains. Within the scope of resilience engineering, it is defined as the "intrinsic ability of a system to adjust its functioning prior to, during or following changes and disturbances, so that it can sustain required operations under both expected and unexpected conditions" (Hollnagel 2011, pp xxxvi). Based on this concept, resilience engineering consists of the development and implementation of the tools necessary to enhance resilience across all system operations (Wreathall, 2006). Although it has been described as an approach to safety management, resilience engineering acknowledges that operations in most industrial sectors rely on highly interdependent sociotechnical systems, and that within such contexts, an effective integration of safety into every aspect of system performance (both at management and operations level) becomes increasingly critical.

As often discussed by Hollnagel *et al* (2006), one of the aims of resilience engineering is the ability to cope with variability of system operations and uncertainty about possible outcomes. Managing variability and uncertainty should be built around four main system capabilities towards resilience (Hollnagel, 2011):

- **Knowing what to do** corresponds to the ability to address the "actual" and respond to regular or irregular disruptions by adjusting function to existing conditions.
- **Knowing what to look for** corresponds to the ability to address the "critical" by monitoring both the system and the environment for what could become a threat in the immediate time frame.

- **Knowing what to expect** corresponds to the ability to address the "potential" longer term threats, anticipate opportunities for changes in the system and identify sources of disruption and pressure and their consequences for system operations.
- **Knowing what has happened** corresponds to the ability to address the "factual" by learning from experiences of both successes and failures.

Enhancing system resilience relies on managing a dynamic balance between these four capabilities. At any given time and place, operational demands may vary considerably and thus system capabilities (resources) must be managed in order to adjust functioning to such changing operational demands.

3 Automation

Automation is defined as when a machine (usually a computer) assumes a task usually performed by humans (Parasuraman & Riley, 1997). Automation is often introduced in order to achieve tasks more efficiently and reliably than humans and the benefits are perceived to be a reduction in human error, a saving on labour costs and a reduction in human workload. However, in complex systems such as rail traffic control, automation lacks the flexibility of human operators in the face of novel situations. Technology is the driving force behind automation, coupled with the desire of organisations to operate systems more economically. However, full automation of complex systems is rare and most automated systems have at least one operator to monitor their performance. Wickens (1992) lists three circumstances when it is appropriate to introduce automation; automation which is employed to perform a function that is beyond the capabilities of a human operator, for example performing complex calculations at high speed; automation which performs functions at which human operators are poor, for example monitoring a system for a single failure event; and automation which provides assistance to human performance, for example augmenting information on display systems.

Issues regarding human interaction with automation are well documented in the human factors literature (e.g. Woods, 1997). Bainbridge (1983) highlighted a number of ironies of automation that included the tendency of designers to automate the tasks that are simple to automate and leave the operator with the tasks that are comparatively more difficult. Thus, automation is often implemented to support operations during normal working, but the operator must take over when conditions move outside the normal operating envelope. The removal of routine tasks from operators may have a resulting effect on their long term skill level (Bainbridge, 1983) and also on their short term situation awareness or a phenomenon associated with automation known as out-of-the-loop unfamiliarity (Endsley, 1996). These effects mean that operators are less equipped to handle disruption efficiently when it occurs.

Balfe et al (2012) discuss the importance of rail traffic control systems providing feedback from the automated system to the operator in order to overcome the issues associated with out-of-the-loop unfamiliarity and Lenoir et al (2006) suggest that feedback may be particularly important in rail traffic control due to the lack of precision in the information available. Kauppi et al (2006) argue that train graphs (a graph

presentation of the timetable) represent a powerful method of presenting information to the operator in a useful format allowing them to proactively identify and manage potential threats.

4 Rail Traffic Control

The purpose of rail traffic control is to ensure separation between trains and to efficiently route trains along the rail network to their destination. For instance, in the British rail network, traffic control is typically achieved through three types of control interface. The first, and oldest, is the lever frame technology dating from the 1800s (Figure 1). These use levers physically attached to the points or signal they control to move the position of the points or signals. The system is operated under the Absolute Block principle, which states that one train may be in one section of the railway on one line at one time. A mechanical interlocking system prevents the signaller from pulling a lever that would set a 'conflicting move', i.e. a route two trains in to the same section of the railway. Signallers communicate the movements of trains by means of a telegraph system which uses bell codes to send messages between signal boxes. This type of control is only suitable for very small areas due to the physical link between the lever and the points/signals.

Fig. 1. Lever frame

The next generation of rail traffic control system, the eNtry-eXit (NX) panel, was introduced in the 1950s (Figure 2). NX panels were introduced in conjunction with technologies to operate points and signals remotely and to provide the signalling system with an indication of train positions. This system allowed signallers to control much larger areas and the development of Track Circuit Block (TCB) also allowed trains to run closer together whilst still ensuring separation, supported by a relay interlocking. The final generation of signalling systems also use TCB and are run similarly to NX panels.

Fig. 2. NX panel

The major change was in the format of the interface, with the control system moving from panel technology to VDU technology (Figure 3). This change facilitated the introduction of automated signalling systems which set the routes for trains according to the timetable. A number of forms of automation are possible, from simple systems which set the priority for trains according to the order in which they appear on the workstation through to advanced forms which attempt to calculate the minimum delay for all trains approaching the area. In the GB network, an advanced form of automation known as Automatic Route Setting (ARS) is used. This system works alongside the signaller and uses complex algorithms to determine which trains to prioritise.

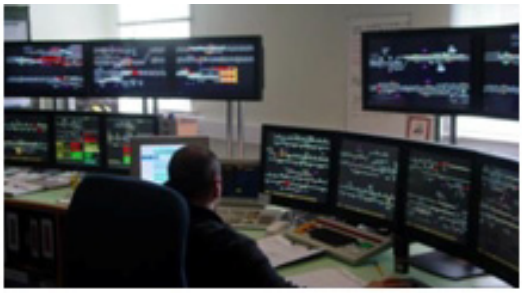


Fig. 3. VDU workstation

In rail traffic control systems, the system works smoothly when running to the predetermined timetable. However, if any disruption occurs due to late running trains or the unavailability of a piece of infrastructure, the signalling task becomes more complex as decisions are required on the routing of trains over a common piece of track. Signallers make a series of decisions on the running order of trains in these circumstances in order to attempt to minimise the effects of the disruption, whilst maintaining safety.

The generation of signalling systems currently under development in GB are Traffic Management systems. These aim to move beyond the traditional control interface of a schematic view of the railway under control towards a train graph (Figure 4) allowing direct manipulation of the timetable. This system can highlight potential

conflicts between trains ahead of time, allowing operators to adjust the timetable to resolve the conflict, in contrast with the current situation where the conflict can only be resolved as it occurs.

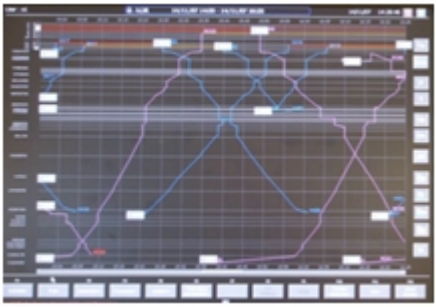


Fig. 4. Train graph

5 Resilience in the Context of Traffic Control

Regardless of the transport sector considered, traffic control is today recognised as a safety critical and complex domain. One of the main repercussions of high complexity is the underspecified nature of systems operations (Wilson *et al* 2009). The large number of human, organisational and technical elements that must be accounted for, together with their fast pace changing behaviour, imposes serious limitations to the ability to fully understand and monitor system operations. Thus, maintaining operational (traffic) control must recognise high variability and uncertainty as constant challenges.

Traffic control is constantly dealing with decisions critical to the safety and efficiency of rail operations as whole. It is inherently built on highly distributed and complex decision making processes. The inevitable finite nature of resources requires the constant making of choices regarding their allocation. At the heart of each of these choices is a decision that must be made. Within the context of rail traffic control, while resource limitation manifests itself through the available access to the rail infrastructure for both the running of trains and the response to any engineering needs, decision making addresses any choices that must be made to safely run as many trains as possible, whilst ensuring the safety of those having to work on the line and reliability of the work they deliver.

Within the scope of resilience engineering, the choices and the decisions that shape them have been expressed through the concept of Efficiency-Thoroughness Trade-Off (ETTO - Hollnagel 2009). In this frame of mind, efficiency generically means achieving a given purpose with minimum expenditure of resources (time, human, technical, financial...) and thoroughness represents a hypothetical ability to accomplish an objective with total disregards to any such resource limitations. It should be kept in mind that actions or devices contributing to safety, because they always consume a certain amount of resources, they constitute a trade-off in favour of thoroughness (although in

the future, they may come to contribute to efficiency in some way). For instance, if nothing else, at least time must be invested to develop a safety check (i.e. pre-flight checklists), and thus, efficiency is sacrificed in some measure.

As resources are always limited, a trade-off between efficiency and thoroughness is inevitably generated throughout every decision making, regardless of the fact that it may often assume a very diffuse nature, rather than an explicit form. Hence, keeping in mind the critical role of decision making earlier described, two capabilities are fundamental for trade-offs to contribute to resilience in traffic control:

- People require information to support their decisions. Safety relies on providing operations and management with information about changing vulnerabilities (Woods & Hollnagel, 2006). Only such information can support an adequate level of awareness regarding how much pressure for efficiency the system can sustain and when it is time to ponder with more thoroughness on the information available, or even to search for additional information (sacrifice decisions).
- Organisations need to develop ways of monitoring safety boundaries. As pointed out by Woods (2006), systems need to maintain awareness and responsiveness to evidence of any potential shifting of decision criteria, which might lead the system across safety limits.

5.1 Analysis of Resilience in Rail Traffic Control Technologies

The four generations of signalling system are analysed here in terms of Hollnagel's (2011) four main system capabilities for resilience of knowing what to do (actual), knowing what to look for (critical), knowing what to expect (potential), and knowing what has happened (factual). The analysis is based on ethnographic observations of signalling operations over a seven-year period as well as a detailed set of interviews conducted with signallers regarding their use and opinions of signalling automation (Balfe et al., 2012). The projections for TM are drawn from experience of TM design philosophies. The outcome of this analysis process is summarised in Table 1.

In terms of the 'actual', the table shows the potential for automation to degrade the signallers' ability to know what to do during disruption situations. The reasons for this are threefold; first, the size of the area controlled by one operator is greatly increased by automation so the decisions required during disruption from one operator are more wide-ranging and complex. Second, as discussed earlier, a long-term consequence of automation is the de-skilling of operators as they are no longer routinely involved in traffic control. Finally, the automation itself will be making decisions and/or suggestions and management pressure to use the costly automation systems mean that the operator must attempt to incorporate the actions of the automation with their own actions.

Table 1. Summary of results drawn from rail traffic control analysis

	Actual	Critical	Potential	Factual
Lever Frame	Small span of control; Limited decision consequences; High degree of control.	Limited span of control; Direct control of all elements; Immediate problems obvious to operator.	Very limited visibility beyond control area; Very limited forewarning of approaching issues.	Factual recording; Opportunity to learn limited to the individual learning from their experiences.
NX Panel	Larger span of control; Larger decision consequences; High degree of control.	Direct control; Large control areas; Problems can go undetected in the short term	Limited visibility beyond control area; Limited forewarning of approaching issues.	Factual recording; Learning is limited to handovers and reports.
ARS	Large span of control; Large decision consequences. Automation may act outside the understood boundaries making it difficult to anticipate.	Indirect control (via automation); Problems can occur without operator knowledge until after the event.	Limited visibility beyond control area; Limited forewarning of approaching issues.	Factual recording; Learning is limited to handovers and reports.
TM	Large and complex span of control; Advanced automation may result in the system becoming beyond the ability of the operator to fully understand and control.	Direct control of the automation via the plan. Visibility of approaching issues via the train graph.	Visibility beyond control area is hugely extended via the train graph.	Potential for replay of events to facilitate organisational learning.

In contrast, the 'critical' dimension has the potential for improvements with the introduction of more advanced automation. This dimension has become progressively more difficult for the signaller to date as the size of their control area has increased (a relatively minor effect) and as comparatively obtuse automation has been introduced. In particular, ARS does not give any indication of its planned actions before imple-

menting those actions meaning the signallers are reliant on their experience to know what the automation will do in a particular situation. However, TM features an improved interface which explicitly displays how the timetable has been modified and gives the key information in a more concise format as well as holding the possibility of highlighting conflicts between trains to operators. This advancement holds the potential to improve the signaller's ability to identify and control threats in the immediate timeframe.

Knowing what to expect in rail traffic control is very dependent on the ability of the operator to see beyond their own area of control in order to identify disruption which may later affect their area. In lever frame boxes, this is extremely limited and the situation is not much improved by the NX and VDU technologies although separate information systems have been implemented to attempt to improve this and the larger area of control means that there is more opportunity to identify issues before they reach the key regulating locations in the area. However, TM does address this by providing operators with visibility of the current and planned train running in adjacent areas and a prediction of the knock-on effects on their own area of control.

Knowing what has happened, 'factual', is not well supported in any of the current systems and organisational learning is haphazard at best. Systems are in place to record the actions taken and the delays to individual trains, but these are primarily used for investigative purposes, not for learning. Again, TM holds the potential for improvement through replay functionality, allowing operators to assess the actions taken for a given scenario and identify more effective solutions. However, this is not core functionality and the efficiency pressures on the organisation may prevent the time and operational staff being made available to take advantage of this opportunity.

6 Discussion

Rail traffic control is a complex system that relies on the decisions of signallers and automated signalling systems to deliver trains to their destination as efficiently as possible, whilst ensuring safety. A key effect of automation has been to increase the area of control of a single operator with consequent effects on their ability to fully understand and correctly control that area. This is compounded by the complexity of the ARS automation system which requires signallers to predict its actions ahead of time in order to fully control it. Although it may be appropriate to automate to allow routes to be set fast and more reliably (Wickens, 1992), some of the pitfalls of automation, such as leaving difficult tasks to the operator (Bainbridge, 1983) have been realised. The result has been the reduction in the overall resilience of the system as failure to correctly control the automation can result in significant additional delays.

In simple terms, although they are still considered traffic controllers, people are more often in the position of overseeing the behaviour of technology, rather than in fact controlling traffic. This contrasts with other domains of control tasks such as air traffic control, where people remain fully responsible for every decision and only rely on technology as a source of information on which to base their decisions. Enhanced feedback from the automated systems is critical for improving overall system

performance and resilience by allowing the automation and the human operator to work more closely together and capitalise on the strengths of each. Access to real time data and information on overall system performance can increase the ability to adjust and adapt to high dynamics operational environments and appropriately designed automation systems can provide this in a format that is accessible and easily processed by the operator.

Traditional automation systems have reinforced rigidity through their basis on the timetable and the comparatively limited approach to prioritisation of trains. Automation towards resilience must support adequate levels of system flexibility and adaptability to cope with dynamics of rail operations and unexpected events, and this can be best achieved by supporting the human operator's decision making process. Currently, railway operations have a very rigid design. They are still based on the principle of blocks (in between two block signals and at any given time, only one train is allowed) and whenever something goes wrong, the system is stopped. From an ETTO perspective, this means that efficiency is always sacrificed in favour of safety. Hence, although the railway can be considered an ultra-safe system, it is not necessarily a resilient one. The integration of new technologies such as those linked to the European Rail Traffic Management System (ERTMS) will introduce new forms of automation and important sources of flexibility into rail systems. Because more flexible operational modes will most likely result in higher degrees of performance variability, in order for such new technologies to contribute to enhanced system resilience, careful consideration must be given to the way in which they will come to support traffic control decision making. The challenge is for the resilience and human factors research community to collect and present clear evidence to guide the design of future traffic control systems and ensure that potential improvements in system resilience are achieved.

On the basis of the analysis presented in this paper, we propose that the attributes of rail traffic automation that can enhance resilience include facilitating a high degree of control over train movements, the ability to anticipate the automation through provision of suitable feedback via intuitive interfaces, increasing the visibility of events in the control area and the boundary areas, and the facilitation of organisational learning through event replay and simulations. The first three attributes address the requirement to provide information to support operator decision-making and the final attribute is proposed as a potential manner in which system safety boundaries can be monitored while also identifying possible improvements in performance.

7 Conclusions

Traffic control is clearly a critical railway component and developments in terms of resilience must take this into account. Thus, increased flexibility in railway systems must be based on the integration of traffic control technology that adequately supports human decision making, which will ultimately be responsible for system ETTOs.

Managing a balance between safety and efficiency under high variability and uncertainty conditions relies on the information available at all hierarchical levels and

organisational areas, and how this information supports decision making with an adequate visibility of operational conditions (Ferreira 2011). As stated by Woods & Hollnagel (2006), progress on safety ultimately depends on providing workers and managers with information about changing vulnerabilities and the ability to develop new means for meeting these.

This paper has discussed resilience only in terms of train routing. Integration of other rail operational domains, including engineering work delivery, electrification, planning, emergency response mechanisms, etc., are not discussed in this paper but are critical to overall levels of system performance. In particular, engineering delivery is not well supported by any generation of signalling system and can be hugely disruptive. This is a key area for future research towards improving resilience in traffic control and in the rail sector as a whole. To this end, further research is needed on rail technology that contributes to a higher integration of various operational domains and needs, aiming to enhance an efficient and safe allocation of critical resources such as access to the rail infrastructure.

References

1. Bainbridge, L.: Ironies of automation. Automatica 19(6), 775–779 (1983)
2. Balfe, N., Wilson, J.R., Sharples, S., Clarke, T.: Development of design principles for automated systems in transport control. Ergonomics 55(1), 37–54 (2012)
3. Endsley, M.R.: Automation and situation awareness. In: Parasuraman, R., Mouloua, M. (eds.) Automation and Human Performance: Theory and Applications, pp. 163–181. Lawrence Erlbaum, Mahwah (1996)
4. Ferreira, P.: Resilience in the planning of rail engineering work. PhD Thesis, University of Nottingham (2011)
5. Hollnagel, E., Woods, D.D., Leveson, N. (eds.): Resilience Engineering – Concepts and Precepts. Ashgate, Aldershot (2006)
6. Hollnagel, E.: Prologue: the scope of resilience engineering. In: Hollnagel, E., Pariès, J., Woods, D., Wreathall, J. (eds.) Resilience Engineering in Practice - A Guidebook, pp. xxix–xxxix. Ashgate, Aldershot (2011)
7. Kauppi, A., Wikstrom, J., Sandblad, B., Andersson, A.W.: Future train traffic control: Control by re-planning. Cognition, Technology & Work 8(1), 50–56 (2006)
8. Lenoir, D., Janssen, W., Neerincx, M., Schreibers, K.: Human-factors engineering for smart transport: Decision support for car drivers and train traffic controllers. Applied Ergonomics 37(4), 479–490 (2006)
9. Leveson, N.: A new accident model for engineering safer systems. Safety Science 42, 237–270 (2004)
10. Muir, B.M.: Trust between humans and automation, and the design of decision aids. International Journal of Man-Machine Studies 27(5-6), 527–539 (1987)
11. Parasuraman, R., Riley, V.: Humans and Automation: Use, misuse, disuse, abuse. Human Factors 39(2), 230–253 (1997)
12. Wickens, C.D.: Engineering Psychology and Human Performance, 2nd edn. Harper Collins, New York (1992)
13. Wickens, C.D., Dixon, S.R.: The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science 8(3), 201–212 (2007)

14. Wilson, J.R., Ryan, B., Schock, A.B., Ferreira, P., Smith, S., Pitsopoulos, J.: Understanding safety and production risks in rail engineering planning and protection. Ergonomics 52(7), 774–790 (2009)
15. Woods, D.D.: Human-centered software agents: Lessons from clumsy automation. In: Flanagan, J., Huang, T., Jones, P., Kasif, S. (eds.) Human Centered Systems: Information, Interactivity and Intelligence, pp. 288–293. National Science Foundation, Washington, DC (1997)
16. Woods, D.: Essential characteristics of resilience. In: Hollnagel, E., Woods, D., Leveson, N. (eds.) Resilience Engineering – Concepts and Precepts, pp. 21–34. Ashgate, Aldershot (2006)
17. Woods, D.D., Hollnagel, E.: Prologue: Resilience Engineering Concepts. In: Hollnagel, E., Woods, D.D., Leveson, N. (eds.) Resilience Engineering – Concepts and Precepts, pp. 1–6. Ashgate, Aldershot (2006)
18. Wreathall, J.: Properties of resilient organisations: An initial view. In: Hollnagel, E., Woods, D.D., Leveson, N. (eds.) Resilience Engineering - Concepts and Precepts, pp. 275–285. Ashgate, Aldershot (2006)

Evaluating Operator's Cognitive Workload in Six-Dimensional Tracking and Control Task within an Integrated Cognitive Architecture

Yan Fu[1], Chunhui Wang[2], Shiqi Li[1], Wei Chen[1], Yu Tian[2], and Zhiqiang Tian[2]

[1] School of Mechanical Science & Engineering,
Huazhong University of Science & Technology, Wuhan, Hubei Province, 430074, China
[2] National Key Laboratory of Human Factors, China Astronaut Training & Research Center,
Beijing, 100094 China
{Laura_fy,sqli}@mail.hust.edu.cn, chunhui89@yahoo.com.cn,
{cctian,tianzhiqi-ang2000}@163.com,
Mileschan@hust.edu.cn

Abstract. Six-dimensional tracking and control task within an Integrated Cognitive Architecture, as a makeup for automated Six-dimensional tracking and control task default. is a common yet highly complex space operation, challenging the human workload. For space exploration system safety, workload is a critical factor in task design and implementation. This research integrates two cognitive architectures: Queuing Network (QN) & Adaptive Control of Thought-Rational (ACT-R) to develop a rigorous computational model for Six-dimensional tracking and control task cognition process. ACT-R represents the human mind as a production rule system. Experiments are set up to build Six-dimensional tracking and control task cognition model and afterwards to validate feasibility of the proposed integrated cognition architecture. Ten subjects of similar training level are chosen to finish manual Six-dimensional tracking and control task with three task difficulty level: one only with displacement margin, one only with posture margin and one with displacement and posture margin. Cognition task analysis is firstly conducted on task performance of subjects. Cognition model of manual Six-dimensional tracking and control task is then built up based on the proposed integration architecture. The proposed integration model developed in the ACTR-QN describes component processes of tracking, decision making and controlling in a 3D environment by ACT-R production rules within QN network. Workload index for each cognition module is calculated based on sector utility throughout the whole task. Human results are compared with the modeled results in the dimension of task time and displacement/posture control trajectory deviation. Workload index is calculated based on the percentage of each module in the time dimension.

Keywords: Mental workload, Simulation, workload, six-dimensional tracking and control task, cognitive modeling.

1 Introduction

Six-dimensional tracking and control task is a very common space exploring task yet highly complex task that involves coordinated control in 3 dimension displacement

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 470–479, 2014.

dimension and 3 dimension posture as well as with the execution of multiple critical subtasks. To explore how astronauts perform this complex task, researchers have developed some models to account for and simulate space driving behavior. Some of these models are primarily conceptual models that help one to understand the representational and procedural components of the driving task[1]. Others are computational models that compute, simulate, and predict various aspects of driving behavior [2-4]. These computational models have emerged as powerful tools for both theoretical study of space driving.

Flight control is the most similar to Six-dimensional tracking and control task. The research com-munity has recently witnessed a growing push for integrated performers models – models that unify the many aspects of flight into a single, larger scale computational model of behavior. Past and ongoing efforts toward integrated flight models, which have shown great promise, accounting for aspects of behavior during air traffic and even performance when flight while performing secondary tasks[5]. But the most popular research in performance modeling lies in road driving behavior. Road driving is 2 dimensional driving. In the case of control mechanism, it is comparable to space driving.

The "artifact" for driving is the vehicle itself and the interface between the human and the vehicle. Embodied cognition is the integrated cognitive, perceptual, and motor processes that manipulate the vehicle and execute the desired tasks[3]. Perception-and-action models of control provide a firm theoretical basis for how perception and action interact in basic tasks such as lateral and longitudinal control [6-8]. The approach to integrated driver modeling explored here centers on the development of driver models in the framework of a cognitive architecture. A cognitive architecture is a general framework for specifying computational behavioral models of human cognitive performance[9-12].The architecture embodies both the abilities and constraints of the human system – for instance, abilities such as memory storage and recall, learning, perception, and motor action; and constraints such as memory decay, foveal versus peripheral visual encoding, and limited motor performance. Anderson proposed ACT-R (Adaptive Control of Thought-Rational) cognitive architecture o model road driving. It is a hybrid architecture based on chunks of declarative knowledge and condition-action production rules that operate on these chunks. Aasman developed a driver model developed in Soar architecture. MHP was proposed to model the air navigation in NASA IMPRINT system[13].

However, Building useful models in ACT-R requires a considerable amount of training and practice. Since ACT-R uses a command-line interface to query the model's internal status, it lacks the visualization of information processing and interactions between its modules. A few efforts have been made to improve the usability of ACT-R as an engineering tool. Previous work, though important, has focused primarily on easier construction of the task knowledge and environment. The research work reported in this paper addresses the visualization issue by representing ACT-R as a Queuing Network (QN), one of whose advantages is the visualization of mental information processing. The QN cognitive architecture has been used to model human performance including reaction time, multitask performance, the psychological refractory period, transcript typing, driving with a secondary in-vehicle

task, and driver workload measured with the NASA-task load index[14]. Such integration is another step towards unified theories of cognition advocated by Allen Newell[15]. We call the integrated architecture in this paper QN-ACTR.

2 QN-ACTR Integrated Cognition Architecture

At the conceptual level, the module network of ACT-R can be represented as a special case of QN. In a QN, information processing is the process of servers holding and processing entities. In ACT-R, modules process information, and buffers hold information. Therefore, modules and their buffers could be considered as servers in QN. Entities flow between these servers and carry the corresponding ACTR information, including buffer requests, chunks, production rules, and the notice of completion that triggers the next service (e.g., the next conflict resolution cycle). The server structure of QN-ACTR is illustrated in Figure 1.

ACT-R represents the human mind as a production rule system. It assumes two types of knowledge representations: declarative chunks and production rules (rules, for short). A chunk's retrieval time and error rate are determined by its activation level, which is jointly determined by the chunk's learning history and association with other chunks. Rules represent procedural knowledge in the form of condition-action (IF-THEN) pairs, and its action will be fired when its condition matches the current "mental state". A mental state consists of the state of each module, and each module is a cognitive component, such as the vision module and the declarative module. ACT-R "thinks" and "acts" by firing rules until a goal state is reached. Figure 1shows the server structure of QN-ACTR. All the servers are ACT-R modules and buffers, and all the paths between servers are information flows in ACT-R.

ACT-R assumes that human has a serial central processor (the production module in ACT-R) and handles multitask scenarios by fast switching between tasks. Each thread represents the task demands from a task. First, it assumes that the goal buffer can hold more than one goal simultaneously. Second, when multiple threads contend for the procedural resource, the least recently processed thread is al-lowed to proceed. Threaded cognition can be incorporated in QN-ACTR as a special case of QN with a specific type of queuing scheduling mechanism. QN-ACTR is implemented in Micro Saint Sharp (http://www.maad.com/), which is chosen because it is a network-based simulation platform and pro-vides natural support for QN modeling.

QN-ACTR was built in a C#-based discrete event simulation software package, Micro Saint Sharp version 2.2. At the implementation level, modules and buffers in ACT-R were programmed as servers (called task nodes in Micro Saint Sharp) as well as the corresponding data objects that store related parameters. Chunks and production rules were programmed as data objects. ACT-R methods and functions were ported to Micro Saint Sharp functions, which can be called by related servers. Global parameters were set to their default values as in ACT-R.

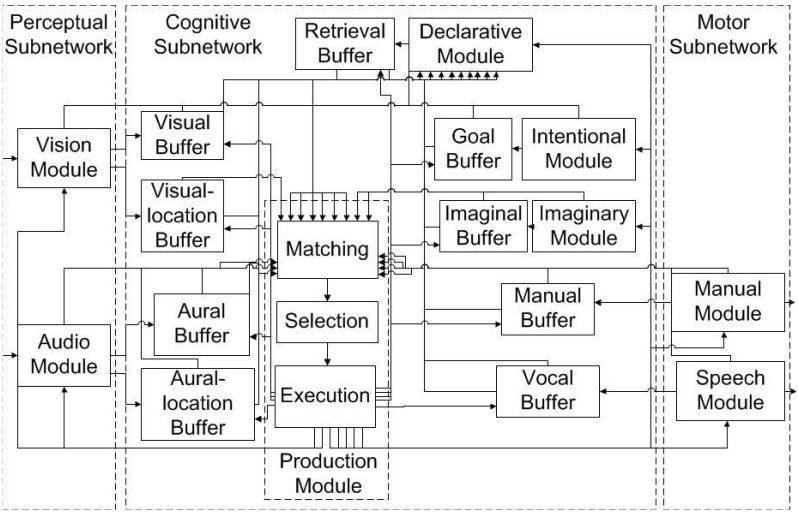


Fig. 1. The server structure of QN-ACTR

The task environment part of a model can be built with task templates supplied with the QN-ACTR. A task template in QN-ACTR is a general description for a type of experiment. A modeler can easily build a task environment by simply setting the template's parameters according to the experiment setup.

After defining the task environment using a template and defining the task-specific knowledge and parameters using the same ACT-R codes, a model is ready to run. In addition to the same text output traces of ACT-R, QN-ACTR can show how information flows in the mind, which is represented and simulated as a QN. For example, Figure 2 is a screenshot that illustrates the implementation of QN-ACTR in Micro Saint Sharp. The server network inside the dashed box represents the same mental structure as the one shown in Figure 1. The server network outside the box represents the task environment (i.e., displays and controls). Servers highlighted by dark borders are busy processing information. In the snapshot of Figure 2, the model is working on three things simultaneously: encoding a visual item, trying to match and select the next production rule, and creating a new chunk in the imaginary module.

QN-ACTR can also visualize the status and details of each module in a separated window. The capability in QN-ACTR can be extended this to audio displays, manual responses, and vocal responses using the "animator" of Micro Saint Sharp. Figure 3 shows a snapshot during the building-sticks task in the ACT-R tutorial.

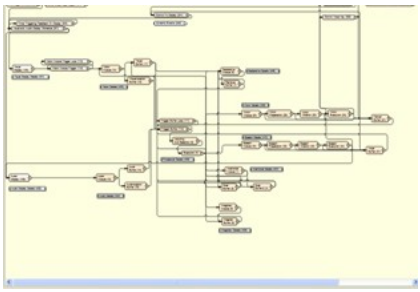


Fig. 2. Visualization of mental information ACTR

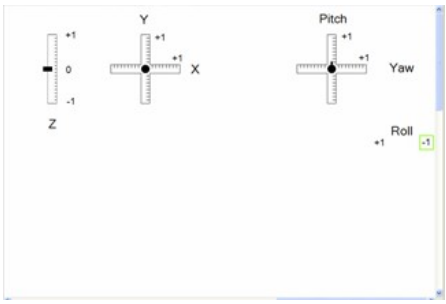


Fig. 3. Visualization of task displays and controls in processing in QN-ACTR

3 Building Six-Dimensional Tracking and Control Task Cognition Model within QN-ACTR Architecture

As mentioned, the ACT-R driver model has three primary components: monitoring, decision-making and controlling. The three components are integrated to run in QN-ACT-R's serial cognitive processor as a tight loop of small cognitive (and related) operations. The entire model is implemented as an ACT-R production system including relevant procedural and declarative knowledge. This section describes each component, the integration of the components into a working implementation, and finally estimation of model parameters and integration with the simulated driving environment.

Control. The control component of the space driver model manages all perception of lower level visual cues and manipulation of spaceship controls for placement control (forward, inward/outward, left/right) and posture control (pitch, yaw, roll).

For simplicity, the model utilizes a longitudinal control law to manipulate forward placement of the ship, very similar to the longitudinal speed acceleration control proposed by Salvucci (2007) for the, namely,

$$\Delta\psi_x = k_{x1}\Delta thw_x + k_{x2} thw_x \Delta t \tag{1}$$

The model encodes the position of the target and derives the time headway thwx to the target. Again, it computes differences from the last instantiation of control, deriving thwx along with the previously mentioned t. These two values then result in an updated value for acceleration.

The acceleration equation attempts to impose two constraints: asteady time headway (Δthwx= 0) and a time headway approximately equal to a desired time headway for following the target. Again, the two constants determine the weights of the two constraints. The acceleration value actually manipulates two controls: A positive value translates to depression of speed acceleration (throttle), and a negative

value translates to depression of speed decrease, with values from 0 to 1 representing no depression to full depression, respectively.

For simplicity, the model utilizes a position difference control law to manipulate forward placement of the ship, very similar to the steering control proposed by Salvucci[16] for the, namely,

$$\Delta\varphi_y = k_{y1}\Delta y + k_{y2}\min(y, y_{\max})\Delta t \tag{2}$$

The model encodes the position of the target and derives the displacement difference Δy to the target. Again, it computes differences from the last instantiation of control, deriving the position difference along with the previously mentioned Δ*z*. These two values then result in an updated value for position difference change. Again, the two constants determine the weights of the two constraints. The value actually manipulates two controls: A positive value translates to depression of difference decrease, and a negative value translates to depression of difference increase, with values from 0 to 1 representing no depression to full depression, respectively.

Posture control is different from speed control and focused on the posture change. For simplicity, the model utilizes a posture control law very similar based on the steering model proposed by Salvucci (2007) for car driving, namely,

$$\Delta\varphi_\theta = k_{\theta 1}\Delta\theta + k_{\theta 2}\min(\theta, \theta_{\max})\Delta t \tag{3}$$

For three direction posture control, the main purpose of the control is to decrease the posture difference between the ship and the target. The paper utilizes the same control law to manipulate the difference in the dimension of pitch, yaw and roll. The control law essentially attempts to impose two constraints: a steady posture degree difference ($\Delta\vartheta = 0$) and a posture degree equal to the maximum degree defined by the task. Again, the two constants determine the weights of the two constraints. The acceleration value actually manipulates two controls: A positive value translates to depression of degree difference increase (throttle), and a negative value translates to depression of degree difference decrease, with values from 0 to 1 representing no depression to full depression, respectively.

Perception. The perception component of the driver model handles the continual maintenance of situation awareness. For this model in the 3-D space environment, situation awareness centers critically on the displacement and posture difference of the spaceship to the target. Perception is currently based on a random-sampling model that checks, with some probability *p*monitor, one of six areas – namely, either forward/backward, inward/outward, left/right, pitch, yaw and roll – with the given decision rules. When the model decides to monitor a particular dimension, it moves visual attention to that dimension and determines whether there is any difference. If so, the model notes the vehicle's current critical dimension in ACT-R's declarative memory. Thus, declarative knowledge continually maintains the awareness of these dimensions. The model could, of course, be extended in a straightforward way to note other 5 dimensions.

Decision-making. The decision-making component of the driver model uses the information gathered during control and monitoring to determine whether any tactical decisions must be made. In the 3-D space environment, the most common decision-making opportunity arises in the determination of which dimension to adjust first and how much is required to adjust.

The decision of *which dimension* to change depends on the possibility of ship moving out of the vision field in certain dimension, given that drivers (in the United States) attempt to stay in the center of matrix originated by the target. If the ship is to move out of the vision field from horizontal dimension, the model checks current difference and time of moving outward. If the difference drops beyond a desired time value, the model decides to change the horizontal dimension

4 Model Validation

Just as no single method, measure, or metric will suffice for understanding human driver behavior, no single one will suffice to validate that the model indeed corresponds well to human driving. Nevertheless, one can validate the most critical parts of a driver model by focusing on key scenarios and analyzing the most important observable data involved in these scenarios. To this end, how the ACT-R model fits several aspects of driver data will now be examined in the scenario Six-dimensional tracking and control task. For this specific scenario, the examination focuses on 6 dimension control output: forward displacement, position difference change in other two dimensions as well as 3 degree changes in pitch, yaw and roll dimensions. The data are compared in the form of aggregate results and time-course profiles.

The computational nature of the QN-ACTR driving model, combined with its ability to interact with the same simulation environment that human drivers use, greatly facilitates the collection and comparison of human and model data. Human data from 10 universities students who are trained well in the simulated scenario. Model data were collected by running ten 10-min model simulations in the same conditions and same environment as the original experiment; note that the model, like a human driver, produces variability in behavior, and thus several simulation runs are desirable to achieve more stable results. The following analysis includes a total of 60 times (20 times at each of 3 difficulty level) of driving data for human participants and same number of times for the model simulations. Because the human and model simulation protocols are identical in form, each set is analyzed in the same manner so as to generate directly comparable measures of driver behavior and performance.

Workload data are sampled based on the statistical function of MicroSaintSharp (See Fig 5). ACTR-QN computed and visualized each sub-network utilization values, which are assumed to have linear relationship with corresponding workload components. Figure 4 shows the utilization of perceptual, cognitive, and motor sub-networks. The visualization clearly demonstrates workload increasing with faster presentation rates and provides more detailed estimation about each workload components.

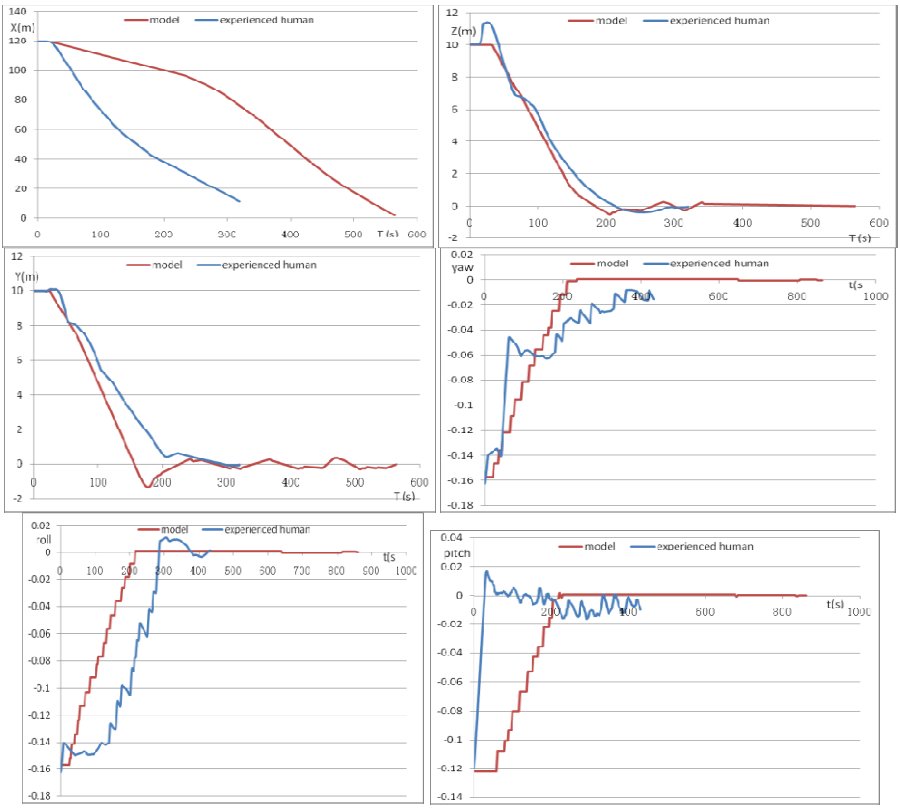


Fig. 4. Tracking Trajectory in 6 dimension (X: forward/backward; Y : left/right; Z: up/down;Pitch, Yaw and roll) for human and model data

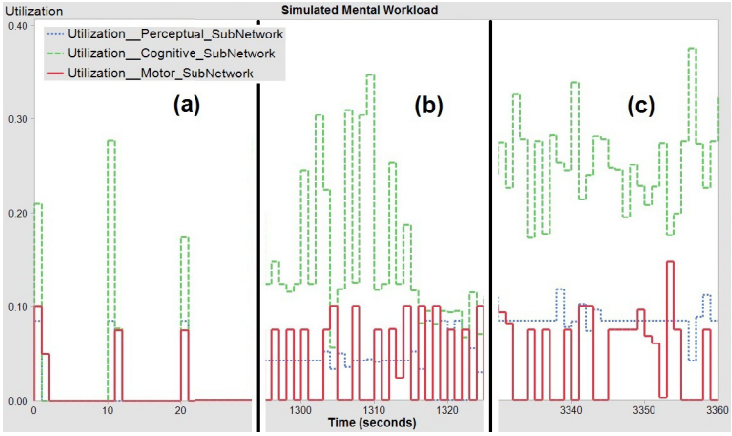


Fig. 5. Visualization of mental workload components in ACTRQNunder different task demands

5 Conclusion

Computational cognitive modeling is quickly maturing to address increasingly complex phenomena at an increasingly high level of rigor. More specifically, cognitive architectures have proven very successful at capturing both lower level performance and higher level decision making in complex dynamic tasks. The QN-ACTR space driving model represents a contribution toward this effort with a novel approach to integrating the lower level (i.e., operational) and higher level (i.e., tactical) aspects of driver behavior in the framework of the QN-ACTR cognitive architecture. Of course, the QN-ACTR Six-dimensional tracking and control task model does not yet provide a complete picture of space driving behavior – further work extending the task, artifact, and/or embodied cognition addressed by the model could take any number of directions. Nevertheless, we are confident that both model and architecture can evolve significantly from the current state of the art to capture a broader and deeper range of the phenomena surrounding driving behavior.

In 3-D space driving, verification results from Six-dimensional tracking and control task model showed that QN-ACTR can produce identical output traces to the human performance (MAPE < 5.0% and R2 > 0.9). The sources of the remaining variances include the difference of built-in random functions between Lisp and C#, which is used in randomly focusing visual attention on the next item, and the difference in rounding digits between Lisp and C#.

QN-ACTR is easy to use. Task-specific knowledge and parameters are defined using the same syntaxes as ACT-R. A task environment is defined by describing the experiment using a task template. The single-discrete-two-stage template is concise and powerful. More templates will be developed to cover other experimental paradigms. Compared with ACT-R, the visualization of the model in QN-ACTR is improved in the aspects of mental information processing and display and control interfaces. Another advantage of QN-ACTR is to define mental workload as network utilization and visualize it. There is currently no theory and measurement for mental workload in the ACT-R 6.0 released version, and the introduction of QN has the potential to improve this.

These mechanisms are what the QN architecture lacks. The QN architecture, on the other hand, represents the mental network with finer granularity. The processing in the QN mental network is more distributed than the processing in ACT-R that centralizes around the procedural module. The procedural module in ACT-R and threaded cognition is assumed to be serial. In contrast, the QN architecture does not have this assumption. Besides, QN does not need executive control to model multitasking performance. We expect that the full integration of ACT-R and QN could combine the advantages from each of them and better model multitasking performance.

In conclusion, QN-ACTR improves the usability of ACTR and the ACT-R implementation of threaded cognition as human factors engineering tools. Future research will examine the benefits of further integration between ACT-R and the QNcognitive architectures.

References

1. Boer, E.R., Hoedemaeker, M.: Modeling driver behavior with different degrees of automation: A hierarchical decision framework of interacting mental models. Paper presented at the 17th European Annual Conference on Human Decision Making and Manual Control held in Valenciennes, France, December 14-16 (1998)
2. Donges, E.: A two-level model of driver steering behavior. Human Factors 20, 691–707 (1978)
3. Groeger, J.A.: Understanding driving: Applying cognitive psychology to a complex everyday task. Psychology Press, Philadelphia (2000)
4. Hildreth, E.C., Beusmans, J.M.H., Boer, E.R., Royden, C.S.: From vision to action: Experiments and models of steering control during driving. Journal of Experimental Psychology: Human Perception and Performance 26, 1106–1132 (2000)
5. Tsimhoni, O., Liu, Y.: Modeling steering using the Queuing Network-Model Human Processor (QN-MHP). In: Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting, pp. 1875–1879. Human Factors and Ergonomics Society, Santa Monica (2003)
6. Fajen, B.R., Warren, W.H.: Behavioral dynamics of steering, obstacle avoidance, and route selection. Journal of Experimental Psychology: Human Perception and Performance 29, 343–362 (2003)
7. Salvucci, D.D.: Inferring driver intent: A case study in lane change detection. In: Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, pp. 2228–2231. Human Factors and Ergonomics Society, Santa Monica (2004)
8. Wilkie, R.M., Wann, J.P.: Controlling steering and judging heading: Retinal flow, visual direction and extra-retinal information. Journal of Experimental Psychology: Human Perception and Performance 29, 363–378 (2003)
9. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111, 1036–1060 (2004)
10. Just, M.A., Carpenter, P.A.: A capacity theory of comprehension: Individual differences in working memory. Psychological Review 99, 122–149 (1992)
11. Levison, W.H., Cramer, N.L.: Description of the integrated driver model (Tech. Rep. No. FHWA-RD-94-092). McLean, VA: Federal Highway Administration (1995)
12. Liu, Y.: Queuing network modeling of elementary mental processes. Psychological Review 103, 116–136 (1996)
13. Aasman, J.: Modeling driver behavior in Soar. KPN Research, Leidschendam (1995)
14. Liu, Y.: Queueing network modeling of humanperformance of concurrent spatial and verbal tasks. IEEETransactions on Systems, Man, Cybernetics 27, 195–207 (1997)
15. Liu, Y., Feyen, R., Tsimhoni, O.: QueueingNetwork-Model Human Processor (QN-MHP): Acomputational architecture for multitask performance inhuman-machine systems. ACM Transactions onComputer-Human Interaction 13(1), 37–70 (2006)
16. Anderson, J.R., Lebiere, C.: The atomic components ofthought. Erlbaum, Mahwah (1998)

Measuring Crew Resource Management: Challenges and Recommendations

Alison Kay, Paul M. Liston, and Sam Cromie

Centre for Innovative Human Systems, School of Psychology, Trinity College, University of Dublin, Ireland
{alison.kay,pliston,sdcromie}@tcd.ie

Abstract. This paper presents a methodology for measuring Crew Resource Management (CRM) parameters as applied to a pilot decision-making task. Six teams of pilots took part in a desk-top decision-making exercise. Flight crew performance was observed by human factors researchers and was measured on a number of parameters pertaining to communication, situational awareness, decision-making, mission analysis, leadership, adaptability and assertiveness. This methodology facilitated the mapping of decisions in the context of the overall process. The communication analysis can be considered more objective than standard CRM expert rating. This methodology could be used to examine CRM for training, recruitment, incident and accident analysis, identifying degraded performance on the flight-deck and has further implications for multi-team co-ordination. It could also be used to provide a sound contribution to the design of automatic means of detection for CRM metrics on the flight deck.

Keywords: teamwork, communication, crew resource management.

1 Introduction

Crew Resource Management was first introduced in the 1980's and has since moved from the world of aviation into other sectors such as healthcare, rail and maritime industries. Good CRM is essential if safe practices are to be upheld regardless of industrial application. CRM research and application in industrial settings have progressed considerably over the last 30 years. Culture changes within organisations over the years and the acceptability of CRM in the workplace has led to CRM being "considered to be a way of working life and it is considered a definitive fact (and is now assumed) that humans do and will make errors and that good CRM is fundamental to recovering from those errors, for managing threats, risks and errors when they present themselves." (Harris, 2011).

2 Challenges for Effective CRM on the Flight-Deck

CRM has been in place in aviation for 30 years and thus is not a new concept. If aviation is such a safe industry, can CRM add anything new to flight deck

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 480–490, 2014.

operations? Could its proliferation create the danger of CRM fatigue? The salient points of CRM have also been transferred into other industries such as healthcare, nuclear, other transport industries, chemical, petrochemical and the process industries with increasing success (Hayward and Lowe, 2010). The civil aviation authority carried out an evaluation of CRM in the UK a number of years ago (CAA, 2003). This evaluation report recommended that the content for single pilot CRM training be examined. This is ever more prescient given that flights may be operated by single crew who are supported from the ground in the not too distant future. How will CRM be affected with these anticipated changes to reduced crew and further increases in automation on the flight deck? Automatic monitoring of CRM could be used to anticipate changes in pilot performance and assist in diagnosing gradually changing levels of incapacitation. Further culture changes in CRM application within organisations are likely if CRM is to be monitored and trained for between ground stations and remotely supported aircraft operators. The research reported herein addresses these questions. Its purpose was to examine CRM metrics as applied to a decision-making task. Human factors researchers analysed CRM parameters within the context of a decision-making task in order to establish whether viewing CRM from multiple angles could give a comprehensive picture of the parameters mapped within the overall process and if this type of picture could then be applied to the operation on the flight deck.

3 Methodology

The validation methodology was based upon previous research which examined distributed situational awareness in a command and control environment (Stewart et al., 2008 and Kay et al., 2008). The methods used were observations, a modified Social Network Analysis (SNA), Hierarchical Task Analysis (HTA), Process Mapping, Co-ordination Demand Analysis (CDA) and Triangulation.

3.1 Observations

Two researchers took part in each data collection phase - both of whom were trained in the collection method for SNA. Researchers positioned themselves near pilots so that they could observe communication. They synchronised their timing devices and made note (using pen and paper) of every instance of communication such as verbal communication, head nod, hand gesture, pointing at the screen. The start and end points for data collection were agreed prior to each data collection session. The raw data from observations was put into electronic format for use in further analyses.

3.2 Modified Social Network Analysis (Communication Counts)

SNA provides a visual and numerical picture of the communication between people. It is used to analyse and represent the peoples' relationships and describes them in terms of how often they communicate, how important people seem to be within a network and how close they may be in the network. This is of interest to this research because it not only gives a visual representation of what the communication is like, but a numerical count of how much communication is taking place. It would not be typical to use SNA for teams of two people as there would generally be a challenge and response nature to the communication (i.e. if one person asks a question, the other person is likely to respond with an answer. Pilots are obliged to communicate in this way on the flight deck). Instead of having the typical network diagram showing multiple people in the network, there would be a figure showing a two people connected by one arrow. This will not provide enough information to make any inference about CRM performance, however, when communication count data is supplemented with information from the process maps a much deeper analysis can be carried out. Being able to comment on the communication frequency between pilots for specific tasks and decision points and being able to determine how information is passed (e.g. verbal commentary, hand signals, written word, and electronic messages) and how this contributes to individual and team contributions to the mission could be invaluable in creating an accurate representation of effective communication and teamwork on the flight-deck.

3.3 Co-ordination Demand Analysis (CDA)

A Hierarchical Task Analysis is carried out as the first step of the CDA. The purpose of the HTA in this research was to provide detailed task information required to feed both the CDA and the process maps. The HTA details the goals and step-by-step tasks involved in a process from start to finish. Each task and subtask within the HTA is classified as either related to task or teamwork. Teamwork related activities are given a rating for each of the teamwork taxonomy criteria. CDA produces a value for the tasks carried out in relation to the total task work, total teamwork and the levels of co-ordination between pilots.

Figure 1 (below) highlights the curricula recommendations for CRM training. The elements listed in both the JAA and FAA recommendations are in keeping with the metrics examined for CDA which are: Communication, Situational Awareness, Decision-making, Mission Analysis, Leadership, Adaptability and Assertiveness (Burke, 2005).

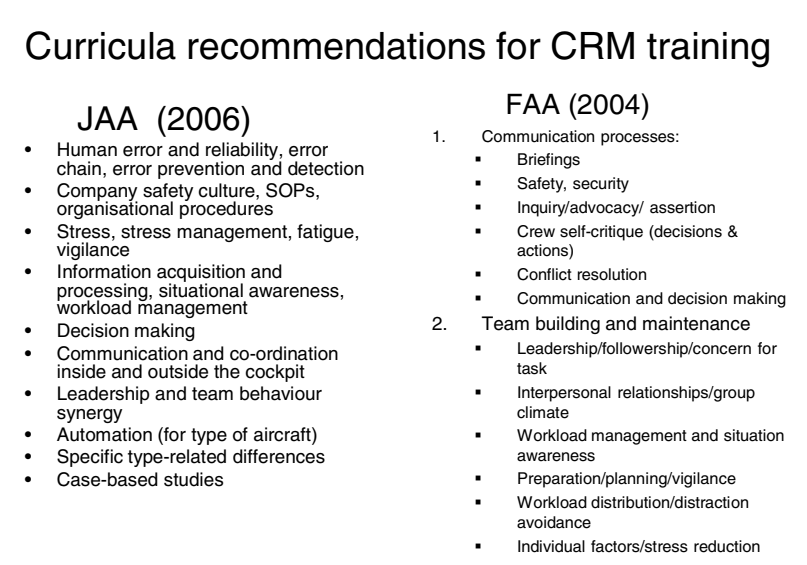


Fig. 1. Curricula Recommendations for CRM training

The commonality of metrics between the curricula recommendations and the current industry standards for measuring CRM and CDA criteria was considered sufficient for CDA to be justifiable as a means with which to examine CRM.

3.4 Process Mapping

A process map for each session was developed. In essence, the process map is a visual representation of the task analysis with a greater level of detail on the people involved in the activities, their resources, flows of information and timelines. The breakdown of tasks and goals that exist within the HTA is mapped out in a systematic manner including the logic and detail of the HTA in conjunction with the visual representation of the communication picture presented in the SNA. This allows researchers to map tasks, subtasks decision points and communication patterns to specific parts of the session. Process mapping affords greater inference of results than an HTA would in linking procedures and processes with individual tasks, personnel, communication and co-ordination. Researchers can then determine how well tasks have been achieved, information has been communicated and how all of this relates to the overall success of the session.

3.5 Triangulation

Researchers can build up a rich picture of each pilot session with SNA – communication counts, decision mapping within the process maps and CDA analysis.

Further triangulation is carried out in the processing of raw data. Transcripts were drawn up and coding applied to all the decision points for each session (i.e. three human factors researchers examined the transcripts and recordings and verified the accuracy of the coding). Coding the transcripts afforded pinpointing of the decision points in the process map and allowed decisions regarding each item being selected to be highlighted to the speech analysts so that they were able to identify which part of the recordings to examine.

3.6 Coding

Recordings from observations are transcribed and coded in order to prepare the data for further analyses using the methods used in this research. The transcriptions are coded to highlight decision points. An example from this research is shown in Figure 2 (below). Each of the 15 items were highlighted a different colour so that they could be easily identified within the text of the transcription. Beside each instance of an item within the transcription, the following reference points and times are noted: 1) the first mention of each item, 2) each decision point made 3) every review of each decision point, 4) the final decision point for each item.

Speaker A: "I think we should definitely take the water (I) (2.17). We don't need a mobile phone (I)(2.18) – there wouldn't be any reception out there anyway. Yes, the water should be first. Do you agree?"

Speaker B: "Absolutely, yes. Water (D1)(2.25) first. The flare (I)(2.26)is pretty important too. What do you think? The radio (I)(2.28)is also important if we hoped to be able to contact people. Is it just a receiver? Maybe we should choose the radio(2.33) first? The water(R1)(2.35) would keep us going for a while, but isn't it more important that people know that we are out there? I think we should go with the radio – would you agree?

Speaker A: "No, I think it just looks like a normal receiver. I don't think that we'd be able to contact anyone using it. I'd say we'd be best going for the water(R2) (2.48) first."

Speaker B: "If it's only a receiver, there's no way it should come before water. Let's go with water(DF)(2.54) first.

Fig. 2. Example of coded transcription

Coding transcriptions enabled researchers to accurately map the decision points into the process map. It also enabled them to examine how decisions are made throughout each session and the impact that external factors (such as the task interference variable) may have had on CRM performance. It would then be possible to see how quickly pilots made decisions, how many times decisions were reviewed and how this related to mission success and the communication frequency outlined in the SNA. These are all fundamental to the overall analyses.

4 Empirical Work

4.1 The Experimental Set-Up

Twelve pilots took part in desktop-based decision-making task which was adapted from The Annual Handbook for Group Facilitators (1975). All pilots were volunteers. Pilots were assigned to teams of two as would naturally occur on the flight deck. They were given a briefing which informed them of the research project, the task that they would be doing, that their speech was going to be recorded and that there would be three observers (2 human factors researchers and 1 speech researcher) in the room. They were also informed that they were taking part on a voluntary basis and could stop at any time. Researchers informed pilots that all data from the trial would be stored safely, made anonymous and that they were free to ask for their data to be withdrawn at any time until the data had been pooled. Pilots were asked to sign consent forms to take part and were given contact details for the researchers should they require further information. They were then shown into the room, (the layout for which is shown in Figure 3)

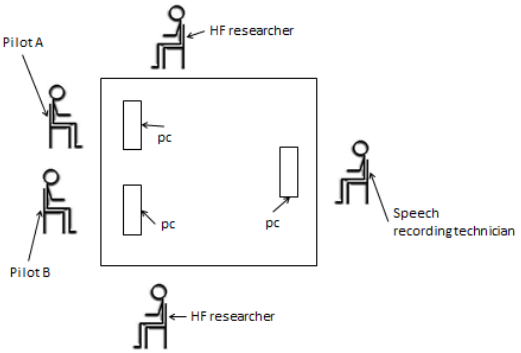


Fig. 3. Diagram of experimental set-up

Pilots had a microphone attached to their lapels and were given headsets through which they were able to hear the "beep" of an incorrect answer. Pilots were given a briefing sheet informing them they are friends who were on a ship that has sunk. Both managed to survive and are now stranded in a life boat in the middle of the sea and are at least 300 miles from the nearest landmass. 15 items were salvaged from the ship before it sank. Pilots were informed that they needed to rank these 15 items in order of importance (starting from the most important item for survival) and told to remember their choices for the session which would last for 10 minutes. All 15 items were displayed on the screen in front of them. Once items had been selected, they remained on the screen. A distractor was introduced to the task. This consisted of a beep in the pilots' headphones which could be interpreted that they had made a

mistake. This beep was generated by one of the researchers. When the 10 minutes were over the pilots were thanked for their time and researchers reiterated that they could be contacted for further information should it be required. Human factors researchers observed the session and kept counts of all types of communication. Speech analysts were present to monitor the recording and to administer the distractor.

4.2 Results

Table 1 (below) shows the pilot group performance for number of items correct and global CDA score ranking.

Table 1. Pilot group performance

Pilot Group	# items	Rank CDA
F	15	1
C	13	2
E	13	3
D	8	4
A	12	5
B	2	6

Pilot group F was the most successful at the decision making task. They also had the highest CDA score. Pilot groups C and E both had the same score for guessing the correct number of items, however Pilot group C had a higher CDA score, thus came 2nd. Pilot Group E came third. Pilot group D came fourth with a higher CDA score than Pilot groups A and B which were 5th and 6th respectively.

Group F had the highest CDA score and the greatest number of items correct. They also had significantly fewer communication counts than all of the other groups and the lowest number of incorrect answers (i.e. beeps). The CDA and decision maps showed that this group also had the lowest number of task steps of the six groups. This group completed the task in 7 minutes whilst all other groups were told that there session was over before they had completed the task.

Group C had the second highest number of items correct and the second highest number of communication. This group experienced 10 “beeps”, most from 8 minutes onwards. The time pressure and distractor had a marked effect on this group’s performance and the last 2 minutes of the session largely consisted of both pilots shouting out random items. There was little continuous dialogue from this point until the end of the session.

Group E also had the second highest number of items correct. Interestingly, this group had the second highest number of “beeps” (16 “beeps”) and the highest number

of communication counts between pilots. Inference could be made that pilots were mitigating for the effect of so many apparent "wrong" answers (i.e. indicated by each "beep"). This would have to be tested as a variable with a larger sample to draw definitive conclusions on this.

Group D had the 4th highest CDA score which may not be reflected in the number of items they got correct (8/15). Their decision-making was very good at the beginning of the session until the distractor "beep" started. This group had the greatest number of "beeps" (22) which seems to have created a great deal of stress. They also had a high communication count which may have mitigated the effect of the distractor. As mentioned for Group E, this would have to be repeated with a larger sample.

Group A came fifth in CDA score. They scored relatively highly (12/15 items) on the number of items correct, however, this group was the most affected by the distractor "beep". There does not seem to be the same mitigating factor of higher levels of communication as in groups C and E. From 7 minutes onwards, pilots seemed to have frozen and there was little or no communication from this point onwards.

Group B had the lowest score for both CDA and the number of items correct (2/15 items). This group displayed excellent prioritisation of items at the beginning of the task. Unfortunately, this was not followed through with decision making. Pilots demonstrated very thorough logic and reasoning of items but did not follow through with decisions on them.

5 Discussion

Without the communication data from the adapted SNA and the ability to examine decisions on the 15 items throughout the whole process of the decision-making task, the results would be lacking considerable detail. The adapted SNA could be considered less subjective in nature than CRM expert rating of communication as it provides a numerical value for the communication that has taken place. There was also more than one rater for each session (thus increasing inter-rater reliability). This data is mapped directly into the overall process and thus provides a more structured framework for further analyses. It could however be argued that the raters were not qualified CRM experts. Criticism levied at non-trained CRM evaluators generally concerns the understanding of the concepts behind the CRM parameters. The raters in this research are experienced human factors researchers with more than 20 years research experience in the aviation industry between them. If any criticism were to be applied here, it could be that they are both non-pilots, however, the task was a non – aviation based one, therefore this criticism is redundant. For future research applying this methodology to the flight-deck, the data will be compared to that of CRM trainers' interpretation of training sessions and will be examined by subject matter experts in incident and accident analyses. This will thus validate the data from an operational perspective.

As mentioned previously, the task used for this research was a non-aviation based one. This had merit for removing the pilots from their flight-deck environment, however, the behavior exhibited may not have been a true reflection of how pilots would have behaved for an aviation-based decision-making task. This is why the next step for the research is apply the methodology directly to behavior on the flight-deck. Interestingly, for each pilot team, the more senior of the pilots took the left hand seat in the room as would be generally represented by the pilot flying or more senior pilot on the flight deck.

The CDA rating scale had a tendency to push the researchers to choose the middle value. If the parameter under scrutiny was not extreme in nature (i.e. '1' or '3'), the researchers were forced to choose '2'. Researchers considered that it would be beneficial for the rating scale to be increased to 5, so that it is possible to give more detail for the performance on the CRM parameter. It would be useful to be able to say that performance was high (5), above average (4), reasonable (3), just under par (2), poor (1). No indication of this was possible using the current rating scale. It is therefore possible that ratings were pushed into the "medium" performance range when it could have been described otherwise. Researchers will adopt this change in rating scale in the next round of research.

The methodological approach from numerous angles make this innovation more powerful in providing a rich picture of CRM mapped into the decision-making task. Internal processes such situational awareness are difficult to assess well. Decision making was somewhat easier to analyse for this research as there was a concrete output attached to decision making in the lost at sea task. These can be clearly identified and mapped into the overall process. Situational Awareness has been shown to be difficult to accurately measure (Kirlik and Strauss, 2006, Pew, 2000, Salmon et al 2006, Stanton et al 2009). This has also been true of the measure of situational awareness in this study which had good face and construct validity but suffered from poor concurrent and predictive validity. Further research to be carried out using this methodology should also accommodate for additional measures such as a freeze technique (e.g. the Situational Awareness Global Assessment Technique) for triangulation specific to situational awareness (as evident in Stanton et al 2009,). Salmon et al 2006 recommend the use of a toolkit in measuring situational awareness. Such toolkits should include 1) performance measures, 2) a freeze probe technique, 3) a post-trial subjective rating scale and 4) an observer rating. This type of toolkit is not unlike the methodology used here.

5.1 Limitations and Further Research

The small sample size in this research has limited the amount of inference that could be made but it is hoped that further studies of simulator sessions on CRM training and live flights will provide a larger data set. The rating scale for CDA is somewhat narrow. There may be a tendency for researchers to choose the middle value. A review of the rating scale will be considered for the next round of analysis. This research did not include parameters for threat and error management. This will also be included for future research.

Contribution to Future CRM Practice. The methodological approach proposed herein and for further research differs from the original HFIDTC study (Stewart et al, Kay et al 2008) in three main areas:1) It encompasses a different systems-approach to task analysis and task modelling. 2) This research will be linked to analyses of CRM criteria from incident data and cockpit voice recordings of selected scenarios available in the public domain. Unfortunately, within the one year lifetime of this research project, but is currently being addressed. 3) In addition to criteria listed in Table 1, Loukopoulos, Dismukes & Barshi (2009) recommend that CRM training be extended for concurrent task management. This will be added as an additional criterion within the teamwork taxonomy. Concurrent task management is a concept which is extremely hard to identify using traditional task modelling and analysis tools. The use of process maps facilitates the identification of concepts such as concurrent task management. This methodology will be applied to simulated flights carried out as part of pilots' CRM training. This data will be compared to the CRM trainers' rating for the sessions. The methodology will also be applied to live flights with multiple teams on board which will facilitate analyses for within and between teams. This work will be evaluated with subject matter experts in CRM training and accident and incident analysis. Thus, the methodology will be used for the analysis of both normal and non-normal operations.

6 Conclusions

Due to the labour intensive nature of this methodology, it is unlikely that it would be employed as standard within organisations, however, it would be of great use in establishing how and where the parameters for CRM fit into the overall process of a flight or mission. This approach would facilitate the design of training for new CRM practices, especially between flight crew in the air and those on the ground for remote ground support. The methods used herein would also be useful in mapping out the gradual decline in performance from task, communication and decision-making perspectives. This would be critical to be able to further identify and define aspects of gradual incapacitation of flight-crew. This research has demonstrated that CRM parameters and decision making can be mapped into the overall process. Incapacitation is generally a gradual process. It is also very difficult to examine incapacitation in applied research settings, thus it would be very useful to be able to be able to map times throughout the flight phase that assistance would be needed from ground support and to be able to back this up using evidence where it was shown that specific CRM metrics suffered at particular points in time. There are many psychophysiological measures taken in measuring pilot performance linked to incapacitation (e.g. galvanic skin response, body temperature, heart rate, eye-tracking). Being able to back these measurements with specific behavioural measures mapped into the overall flight operational process is fundamental to being able to understand a rich picture of what gradual incapacitation looks like if there it is hoped that we will be able to automatically detect it happening on flight decks in the future. The methodology used herein could support such analyses.

In conclusion, this methodology examines CRM along similar parameters to those of LOSA and LOFT (with the exception of threat and error management), but also includes more objective measures for communication analyses. The CRM parameters and decision points can be mapped into the overall operations process, so that patterns and clusters of communication and activity can be examined in greater detail. This research has been innovative in its approach to the measurement of CRM metrics. There has been considerable effort to examine CRM metrics from several angles (SNA, CDA, Process Mapping) and the contribution of more objective measurement (communication) has been invaluable.

References

1. Burke, S.C.: Team Task Analysis. In: Stanton, N.A., Salmon, P.M., Walker, G.H., Baber, C., Jenkins, D.P. (eds.) Human Factors Methods: A Practical Guide for Engineering and Design, Ashgate, Hampshire, UK, pp. 56.1 – 56.8 (2005)
2. CAA, Crew Resource Management (CRM) Training: Guidance for flight crew, CRM instructors (CRMIS) and CRM Instructor Examiners (CRMIES) (CAP 737). Civil Aviation Authority, London (2006)
3. CAA, Methods used to Evaluate the Effectiveness of Flightcrew CRM Training in the UK Aviation Industry, CAA Paper 2002/05. Civil Aviation Authority, London (2003)
4. Flin, R., O'Connor, P., Crichton, M.: Safety at the Sharp End: A guide to non-technical skills, Ashgate, Surrey, UK (2008)
5. Harris, D.: Human Performance on the Flight Deck, Ashgate, Surrey, UK (2011)
6. Hayward, B.J., Lowe, A.R.: The Migration of Crew Resource Management Training. In: Kanki, B., Helmreich, R., Anca, J, eds. (2010) Crew Resource Management, ch. 12, 2nd edn., Wiley, San Diego (2010)
7. Kanki, B., Helmreich, R., Anca, J.: Crew Resource Management, 2nd edn. Wiley, San Diego (2010); Kay, A., Lowe, M., Salmon, P.S., Stewart, R., Tatlock, K., Wells, L.: Case Study in RAF Boeing E3D Sentry. In: Stanton, N.A., Baber, C., Harris, D, eds. Modelling command and Control, Ashgate, Surrey (2008)
8. Kirlik, A., Strauss, R.: Situation awareness as judgment I: Statistical modeling and quantitative measurement International Journal of Industrial Ergonomics 36 (2006)
9. Pew, R.W.: The state of situation awareness measurement: Heading toward the next century. In: Endsley, M.R., Garland, D.J. (eds.) Situation Awareness Analysis and Measurement, pp. 33–50. Erlbaum, Mahwah (2000)
10. Salmon, P.M., Stanton, N.A., Walker, G.H., Jenkins, D., Ladva, D., Rafferty, L., Young, M.: Measuring situation awareness in complex systems: comparison of measures study. International Journal of Industrial Ergonomics 39(3), 490–500 (2009)
11. Salmon, P.M., Stanton, N.A., Walker, G., Green, D.: Situation awareness measurement: A review of applicability for C4i environments. Applied Ergonomics 37, 225–238 (2006)
12. Stanton, N.A., Baber, C., Harris, D.: Modelling command and Control: Event Analysis of Systemic Teamwork, Ashgate, Surrey, UK (2008)
13. Stewart, R.J., Stanton, N.A., Harris, D., Baber, C., Salmon, P., Mock, M., Tatlock, K., Wells, L., Kay, A.: Distributed situational awareness in an airborne warning and control aircraft: application of a novel ergonomics methodology. Cognition Technology and Work (10), 221–229 (2008)
14. Pfeiffer, J.W., Jones, J.E.: The 1975 Annual Handbook for Group Facilitators. University Associates, Incorporated, La Jolla (1975)

Study on Diagnosis Error Assessment of Operators in Nuclear Power Plants

Ar Ryum Kim[1], Inseok Jang[1], Jaewhan Kim[2], and Poong Hyun Seong[1]

[1] Department of Nuclear and Quantum Engineering, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, Republic of Korea
{arryum,nuclear82,phseong}@kaist.ac.kr
[2] Integrated Safety Assessment Division, Korea Atomic Energy Research Institute, 450-1 Dukjin-dong, Yuseong-gu, Daejeon, 305-353, Republic of Korea
jhkim4@kaeri.re.kr

Abstract. The purpose of this study is to suggest a framework to assess diagnosis error of operators in nuclear power plants. In nuclear power plants, human error caused by inappropriate performance due to inadequate diagnosis of situation by operators have been considered to be critical since it may lead serious problems. In order to identify and estimate the human errors, various human error analysis methods were developed so far. Most human error analysis methods estimate diagnosis error through time reliability curve or expert judgments. In this study, a new framework to assess diagnosis error was suggested. It is assumed that diagnosis error is caused by inadequate quality of data and diagnosis error can be observed by using information processing model of human operators. Based on this assumption, we derived the assessment items for the quality of data and diagnosis error taxonomy here.

Keywords: Diagnosis errors, Quality of data, Information processing model.

1 Introduction

Diagnosis errors which may cause inadequate actions of operators have been considered to be critical since it may lead fatal problems in the safety critical systems such as chemical plants, airlines and nuclear power plants (NPPs). In NPPs, when operators experience the changing situations of the plants, they may diagnosis the situations. They will interpret the cause of problems, observe the relevant cues, plan the tasks to resolve the situation and expect the consequences. As a result of this process, they will make the decisions and implement the proper actions. In this regards, when operators fail to diagnose the situation correctly, it will lead inadequate implementations of actions and may make the plant in danger.

In order to identify and estimate human errors, various human reliability analysis (HRA) methods have been developed so far. Most HRA methods assess diagnosis errors through time reliability curve (TRC) and expert judgment as shown in Table 1. Thus most HRA methods believe that if operators have sufficient time available to

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 491–498, 2014.

diagnose, the diagnosis error may be decreased dramatically. However, diagnosis error can be affected by a diagnostic ambiguity that may arise from various plant responses [1] and the quality of information.

The purpose of the study is to propose the framework to assess diagnosis errors of the operators in NPP main control room (MCR). For the first step of the study, we identified the contributors to diagnosis errors and developed diagnosis error taxonomy.

Table 1. Approaches to estimate diagnosis errors in the existing HRA methods

	TRC	Expert judgment
The existing HRA methods	-THERP (Swain, 1987) - HCR (EPRI, 1992) - K-HRA (KAERI, 2005) - SHARP (EPRI, 1984)	- INTENT (INL, 1992) - HRMS (B. Kirwan, 1997) - HEART (J.C. Willliam, 1988) - SLIM (BNL, 1988) - SPAR-H (INL, 1995)

2 A Framework to Assess Diagnosis Errors

D. I. Gertman asserted that the cognitive errors stem from erroneous decision making, poor understanding of rules and procedures, and inadequate problem solving and this errors may be due to the quality of data and people's model for processing information [2]. E. Hollnagel said that information processing models of human behavior assume that there are reliable criteria of validity against which it is possible to measure a deviant response [3].

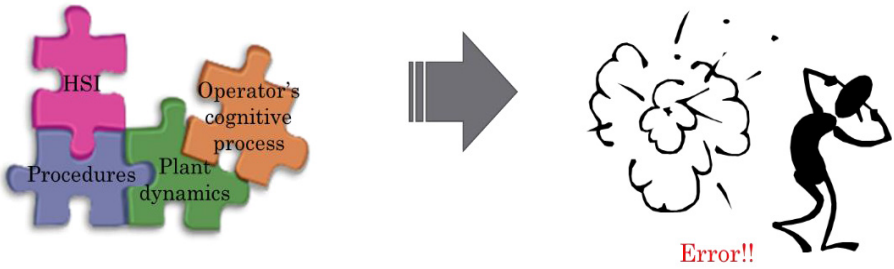


Fig. 1. The contributors to diagnosis errors of operators

As shown in Figure 1, we assume diagnosis error can be caused due to the quality of information (inadequate information from human-system interface (HSI) and procedures, the erroneous situations from other causes) and diagnosis error can be observed by people's model of processing information (not qualified cognitive process of operators). Here, we assume that information of HSI and procedure are the contributors to diagnosis error.

In order to derive assessment items for the quality of data, literature review was performed. For developing diagnosis error taxonomy, naturalistic decision making (NDM) and predictive human error analysis (PHEA) were used. For validation of the assessment items and diagnosis error taxonomy, audio-visual recorded data under training session by operators and NPP accident reports from 2010 to 2013 have been analyzed.

2.1 Identification of Assessment Items for Quality of Data

Literature review have been performed in order to derive the assessment items for the quality of data. Twelve papers related to diagnosis errors were reviewed in nuclear, aviation, and medical domains from 1988 up to now [4-15]. In addition, six general HRA methods were reviewed [16-20]. Most papers asserted that improper information such as inadequate HSI design and poor procedures development, too many alarms and multiple concurrent events are one of major contributors to diagnosis errors. Also, dynamic situation of systems combined with a specific initiator at an earlier time of an event scenario may increase diagnosis errors of operators [1]. We categorized assessment items into three parts: HSI information, procedures and others causes. As a result, eight assessment items for HSI information, six items for procedures, and five items for other reasons have been derived as shown in Table 2.

However, the assessment items derived should be confirmed and validated to whether or not those are suitable to the unique situation of NPPs. For the validation, training data which have been recorded audio-visually and NPP accident reports have been analyzed.

Table 2. Assessment items for HSI information, procedures and other causes

	HSI information	Procedures	Other causes
Assessment items	-The number of alarms alerted (Single/Multiple) -Location (MCR/local) -Directedness -Indicated in procedures or not -Training of HSI -Accuracy of information	-Existence -Including a seperated check sheet or not -AND or NOT wording in the text -Training of procedures -The number of procedures used (Single/Multiple) -Standard or ambiguous working	-Temporal characteristics -Multiple events -Delayed system response - Time available - Training - ...

2.2 Development of Diagnosis Error Taxonomy

In order to develop diagnosis error taxonomy, information processing model of human operators is used. Because information processing models of human behavior assume that there are reliable criteria of validity against which it is possible to measure a deviant response [3]. Among vaious models, naturalisitc decision making (NDM) is selected.

NDM was developed to describe how people actually make decisions as a result of cognitive processing in real-world settings [21]. The study of NDM asks how experienced people, working as individuals or groups in dynamic, uncertain, and often fast-paced environments, identify and assess their situation, make decisions and task actions whose consequences are meaningful to them and the larger organization in which they operate [22]. Here, it is assumed that people does not compare options parallerly. When people meet changing situation, they rapidly search thier experience which is most similiar to the current situation. Based on the prior experience, they suggests the several solutions and evaluate solutions serially. NDM researchers have studied people in filed settings, such as nuclear power plants, navy commanders, anesthesiologist, and airline pilots [23]. F.L. Greitzer [24] and P. Carvalho [25] also examined the cognitive process of the expert operators in NPPs when they make critical decisions and asserted that NDM is suitable.

Recognition primed decision making (RPD) model is most widely used model of NDM [26]. This model describes how people use their experience in the form of a repetoire of patterns. When people need to make a decision, they can quickly assess the situation by matching petterns they have learned. If they find a clear match, they can evaluate what is the most typical course of action by mental simulaton [23]. Thus, RPD model incorporates two cognitive processes: 1) Assess the situation by pettern matching and 2) Evaluate course of action by mental simulation. The detailed cognitive process of RDP model is shown in Figure 2.

In order to develop diganosis error taxonomy, we adopt the human error identification (HEI) techniques. Among various HEI techniques, we selected predictive human error analysis (PHEA).

As a results, we derive eleven diagnosis error modes: cues are not observed, wrong cues are observed, insufficent cues are observed, goal preconditions ared ignored, incorrect goal is set, correct but insufficient goal is set, wrong picture (expectation) is derived, wrong option is generatd, acton is evaluated incorrectly, correct action is generated but too soon/too late, and correct action is generated but wrong order.

However, the diagnosis error taxonomy should be also confirmed and validated to whether or not those are suitable to the unique situation of NPPs. For the validation, training data which have been recorded audio-visually and NPP accident reports have been analyzed.

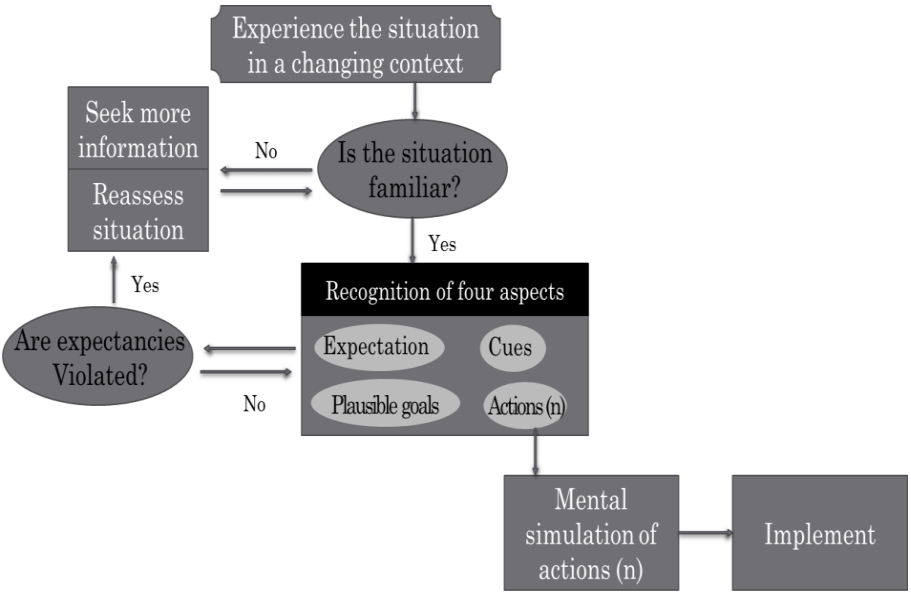


Fig. 2. The detailed cognitive process of RPD model

3 Case Study

Here, for the case study, simulation data from training session by operating teams were selected. During training session, all communications between operatros were recorded in an audio-visual format and transcribed to the verbal protocol data.

Among many scenarios during training sessions, two scenarios were selected as shown in Table 3. Three same operating teams were participated in both two scenarios. During both scenarios, operators should perform the tasks which are addressed in standard post trip action (SPTA), diagnostic action (DA), and emergency operating procedures (EOP) procedures after the reactor trip.

Table 3. Scenario descriptions

	#1	#2
Scenario	Steam generator tube rupture (SGTR) + Safety injection (SI) actuation failure	Station black out (SBO)
# of operating teams	3	3
Procedures (after reactor trip)	SPTA/DA/EOP-03	SPTA/DA/EOP-07

In the scenario #1, during performing SPTA procedures, the signals for SI actuation failure was occurred. Then, operators should follow the SPTA procedures and diagnose the situation to resolve SI actuation failure concurrently. A result of the case study is shown in Table 4.

Table 4. The result of case study

	#1	#2
HSI information	-Multiple alarms are alerted. -MCR information is necessary. -All indicators are addressed in procedures. -All information are accurate.	-Multiple alarms are alerted. -MCR information is necessary. -All indicators are addressed in procedures. -All information are accurate.
Procedures	-Procedure is existed. -Procedure does not include separate sheet. -Procedure includes AND logic.	-Procedure is existed. -Procedure does not include separate sheet. -Procedure includes AND logic.
Others	-**Multiple events** are occurred.	-**Single event** is occurred.
Diagnosis errors	-2/3 teams failed to diagnose correctly. Team 1: Wrong cues were observed. Team 3: Cues were not observed.	-No team was failed.

As shown in Table 4, all other information for HSI and procedures were same for both scenarios except plant dynamics: Multiple and signle event. However, these difference caused the significant increase of diagnosis errors. In the case of scenario #1, two out of three teams failed to diagnose the SI actuation failure. Only one team diagnosed the situation correctly and resolved the problem. While, in the case of scenario #2, no team failed diagnosis during performing SPTA procedures. In this regards, it is confirmed and validated that multiple event leads diagnosis error obviously.

4 Summary and Conclusion

Diagnosis errors may cause inadequate actions and lead the unwanted situation of nuclear power plants. It is crucial to identify and estimate diagnosis error to assure the safety of nuclear power plants. However, most HRA methods assess diagnosis error through time reliability curve or expert judgments. Also, these HRA methods does not consider the situation such as inadequate information is represented in HSI or procedures. In order to incorporate the assessment for quality of information (HSI,

procedure, and plant dynamics), the research have been conducted here. For the first step, we derived the assessment items for the quality of data through literature review.

As a results, eight assessment items for HSI information, seven assessment items for procedures, and five assessment items for the other causes were derived. In order to derive diagnosis error taxonomy, recognition primed decision-making model which is most widely used model for naturalistic decision making and predictive human error analysis which is the one of human error identification technique were used. As a result, eleven diagnosis errors were derived.

For confirmation and validation of assessment items and diagnosis error taxonomy, training data and accidents reports have been analyzed. As a case study, two scenarios were compared to validate assessment items for quality of data. This case study represented that multiple event may cause the significant increase of diagnosis errors. Thus, the possibility of multiple events should be considered to estimate diagnosis errors.

References

1. Kim, J.W., et al.: The MDTA-based method for assessing diagnosis failures and their risk impacts in nuclear power plants. Reliability Engineering and System Safety 93, 337–349 (2008)
2. Gertman, D.I., et al.: INTENT: a method for estimating human failure probabilities for decision based failures. Reliability Engineering and System Safety 35, 127–136 (1992)
3. Hollnage, E.: Cognitive reliability and error analysis method (1998)
4. Kun, M., et al.: A dynamic neural network aggregation model for transient diagnosis in nuclear power plants. Progress in Nuclear Energy 49(3), 262–272 (2007)
5. Kang, H.G., et al.: Application of condition-based HRA method for a manual actuation of the safety features in a nuclear power plant. Reliability Engineering and System Safety 91(6), 627–633 (2006)
6. Kun, M., et al.: A neural network based on operation guidance system for procedure presentation and operation validation in nuclear power plants. Annals of Nuclear Energy 34(10), 813–823 (2007)
7. Kim, M.C., et al.: A computational method for probabilistic safety assessment of I&C systems and human operators in nuclear power plants. Reliability and System Safety 91(5), 590–593 (2006)
8. Jung, W.D., et al.: Analysis of operators' performance time and its application to a human reliability analysis in nuclear power plants. IEEE Transactions on Nuclear Science 54(5), 1810–1811 (2007)
9. Jambon, F.: Taxonomy for human error and system fault recovery from the engineering perspective. In: International Conference on Human-Computer Interaction in Aeronautics, Montreal, Canada, May 27-29, pp. 55–60 (1998)
10. Reer, B.: Evaluation of new developments in cognitive error modeling and quantification: time reliability correlation. Probabilistic Safety Assessment and Management 96, 645–650 (1996)
11. Le Max-product, T.: algorithms for the generalized multiple-fault diagnosis problem. IEEE Trans. Syst. Man Cybern. B Cybern. 37(6), 1607–1621 (2007)
12. Orasanu, J., et al.: Errors in aviation decision making: A factor in accidents and incidents. In: HESSD 1996, pp. 100–107 (1996)

13. Canosa, R.L., et al.: Modeling decision-making in single-and multi-modal medical images. In: SPIE Proceedings, vol. 7263 (2009)
14. Park, J.K., et al.: Identifying cognitive complexity factors affecting the complexity of procedural steps in emergency operating procedures of a nuclear power plants. Reliability Engineering and System Safety 89(2), 121–136 (2005)
15. Ujita, H., et al.: Development and verification of a plant navigation system displaying symptom based procedure. Cognition, Technology & Work 3(1), 22–32 (2001)
16. Kirwan, B.: The development of a nuclear chemical plant human reliability management approach: HRMS and JHEDI. Reliability Engineering and System Safety 56, 107–133 (1997)
17. Chien, S.H., et al.: Quantification of human error rates using a SLIM-based approach, Human factors and power plants, Monterey, CA, USA, May 5-9, pp. 297–302 (1988)
18. U.S. NRC: The SPAR-H Human reliability, analysis method, NUREG/CR-6883 (2005)
19. KAERI: Development of a standard method for human reliability analysis (HRA) of nuclear power plants –Level 1 PSA full power internal PSA, KAERI/TR-1961/2005 (2005)
20. EPRI: An approach to the analysis of operator actions in probabilistic risk assessment, EPRI TR-100259 (1992)
21. Klein, G.: Naturalistic decision-making. Human Factors 50(3), 456–460 (2008)
22. Zsambok, C.E.: Naturalistic deicision making. Lawrence Erlbaum Associates, Mahwah (1997)
23. Klein, G., et al.: Naturalistic decision making. Human System IAC 2(1), 16–19 (1991)
24. Graitzer, F.L., et al.: Naturalistic decision making for power system operators. International Journal of Human Computer Interaction 26(2-3), 278–291 (2010)
25. Carvalho, P., et al.: Nuclear power plant shift supervisor's decision making during microincidents. International Journal of Industrial Ergonomics 35, 615–644 (2005)
26. Elliott, T.: Expert decision-making in naturalistic environments: a summary of research, Defend science and technology organization, DSTO-GD-0429 (2005)

Task Switching and Single vs. Multiple Alarms for Supervisory Control of Multiple Robots

Michael Lewis[1,*], Shi-Yi Chien[1], Siddarth Mehortra[2], Nilanjan Chakraborty[2], and Katia Sycara[2]

[1] University of Pittsburgh, School of Information Sciences, Pittsburgh, PA 15260, USA
ml@sis.pitt.edu, gsechien@gmail.com
[2] Carnegie Mellon University, Robotics Institute, Pittsburgh, PA 15213, USA
siddarthmehrotra11@gmail.com, {nilanjan,katia}@cs.cmu.edu

Abstract. Foraging tasks, such as search and rescue or reconnaissance, in which UVs are either relatively sparse and unlikely to interfere with one another or employ automated path planning, form a broad class of applications in which multiple robots can be controlled sequentially in a round-robin fashion. Such human-robot systems can be described as a queuing system in which the human acts as a server while robots presenting requests for service are the jobs. The possibility of improving system performance through well-known scheduling techniques is an immediate consequence. Unfortunately, real human-multirobot systems are more complex often requiring operator monitoring and other ancillary tasks. Improving performance through scheduling (jobs) under these conditions requires minimizing the effort expended monitoring and directing the operator's attention to the robot offering the most gain. Two experiments investigating scheduling interventions are described. The first compared a system in which all anomalous robots were alarmed (Open-queue), one in which alarms were presented singly in the order in which they arrived (FIFO) and a Control condition without alarms. The second experiment employed failures of varying difficulty supporting an optimal shortest job first (SJF) policy. SJF, FIFO, and Open-queue conditions were compared. In both experiments performance in directed attention conditions was poorer than predicted. A possible explanation based on effects of volition in task switching is proposed.

Keywords: human-robot interaction, neglect tolerance model, scheduling, task-switching.

1 Introduction

In the simplest case of multirobot control, an operator controls multiple independent robots interacting with each as needed. A foraging task [1] in which each robot searches its own region would be of this category. Control performance at such tasks can be characterized by the average demand of each robot on human

* Corresponding author.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 499–510, 2014.

attention [2]. Such operator interactions with a robot might be described as a sequence of control episodes in which an operator interacts with the robot for period of time (interaction time, IT) raising its performance above some upper threshold (UT) after which the robot is neglected for a period of time (neglect time, NT) until its performance deteriorates below a lower threshold (LT) when the operator must again interact with it. In practice the operator's task is even more complex. Humans are additionally included in robotic systems to perform tasks the automation cannot. The most common of these tasks is searching for targets in noisy displays such as remote video or aerial imagery.

Research in robot self-reflection [3] has progressed to the point that it is plausible to presume robots capable of reporting their own off normal conditions such as an inability to move or unsafe attitude. By focusing the operator's attention on robots needing interaction rather than requiring the operator to monitor for the failures, time spent monitoring can be eliminated increasing the number of robots that can be serviced over the intervening interval. With robots informing the operator of their need for interaction the human-robot system becomes more like a queuing system in which the operator acts as the server and robot interaction requests as jobs. Using operations research methods the performance of such a queuing system might be further improved by prioritization of jobs or adjustment of service levels [4] to match current conditions. Deriving full benefit from such aiding, however, would require the ability to focus an operator's attention on a particular robot. We refer to the possibility that human attention might be closely directed in this manner without loss of cognitive efficiency as the attention scheduling hypothesis.

Alarms are commonly used in complex human-machine systems to direct human attention but usually in an open and unrestrictive way. Annunciator systems in nuclear power plants or aircraft cockpits typically alarm separately for each setpoint that has been exceeded allowing the human to prioritize and schedule attention among competing demands. Human-multirobot tasks exert similar competing demands on operators frequently requiring them to mix navigation, visual search, and status monitoring to accomplish their objectives. If operators can manage their own attentional resources to avoid damaging interruptions and/or exploit common situational elements among tasks these advantages might outweigh benefits available from externally directed attention.

Experiment I tests the attention scheduling hypothesis by comparing operators performing a multirobot foraging task without alarms for robot failures, with all alarms available (Open-queue), or with a first-in-first-out (FIFO) queue making only a single alarm available at a time. Effects were measured for both the primary task of searching for and identifying victims and the secondary task of identifying and restoring failed robots. Because all failures were of the same difficulty, the order in which they were serviced should make no difference so under the attention scheduling hypothesis the FIFO and open alarm conditions should produce equivalent performance.

Experiment II extends the test to a condition under which the attention scheduling hypothesis would predict superior performance for directed attention. The shortest job first (SJF) discipline is a provably optimal policy for maximizing throughput in a

queuing system [5]. Using this policy to direct human attention, therefore, should lead to superior performance under the attention scheduling hypothesis providing the undirected operators did not follow precisely the same policy. This experiment compares SFJ, FIFO, and Open-queue conditions with attention scheduling hypothesis predictions that SJF should produce the best performance followed by Open-queue provided that operators did better than random (FIFO) in selecting robots to be serviced.

2 Methods

The reported experiments were conducted using the USARSim robotic simulation with simulated Pioneer P3-AT robots performing an Urban Search and Rescue (USAR) foraging task. USARSim is a high-fidelity simulation of USAR robots and environments developed as a research tool for the study of human-robot interaction (HRI) and multi-robot coordination. USARSim supports HRI by accurately rendering user interface elements (particularly camera video), accurately representing robot automation and behavior, and accurately representing the remote environment that links the operator's awareness with the robot's behaviors. USARSim uses Epic Games' UnrealEngine3 to provide a high fidelity simulator at low cost and also serves as the basis for the Virtual Robots Competition of the RoboCup Rescue League. Other sensors including sonar and audio are also accurately modeled.

MrCS (Multi-robot Control System), a multi-robot communications and control infrastructure with accompanying user interface, developed for experiments in multirobot control and RoboCup competition [6] was used in these experiments. MrCS provides facilities for starting and controlling robots in the simulation, displaying multiple camera and laser output, and supporting inter-robot communication.

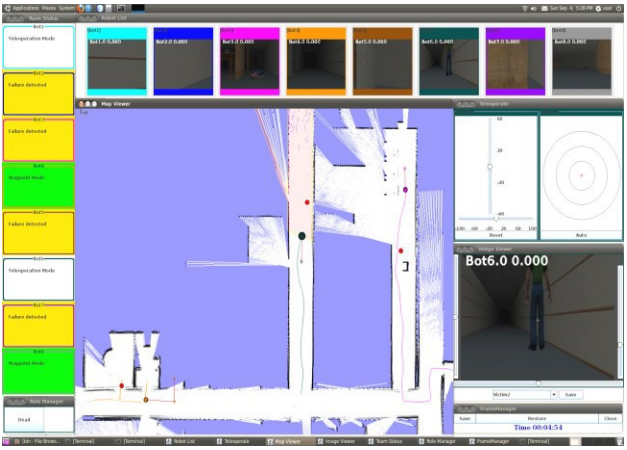


Fig. 1. MrCS Open-queue condition with status bar on left

Figure 1 shows the MrCS user interface in the Open-queue alarm condition. Thumbnails of robot camera feeds are shown on the top, the currently selected video feed of interest in the bottom right. A GUI element in the middle right allows teleoperation and camera pan and tilt. Current locations and paths of the robots are shown on the Map Viewer (middle) which also allows operators to mark victims. The team status window (left) for the Open-queue condition shows each robot's current status and briefly summarizes any problem. Green indicates the robot is in autonomous condition and functioning safely, yellow indicates an abnormal condition, such as stuck at a corner. When a robot is manually controlled, its tile turns white. The operator selects the robot to be controlled from either the team status window or camera thumbnail.

In forced queue conditions robots in abnormal states are presented one at a time. Additional alarms can only be reviewed after the presenting problem is resolved. To avoid "clogging" the status window with an unrecoverable failure, operators have an alternative in a "Dead" button. Once switched off, the robot will stop reporting and no longer be scheduled. The status panel is removed in the Control condition requiring operators to monitor the Map Viewer and thumbnails to identify malfunctioning robots.

When an operator detects a victim in a thumbnail, a complex sequence of actions is initiated. The operator first needs to identify the robot and select it to see the camera view in a larger window and to gain the ability to stop or teleoperate the robot. After the user has successfully selected a robot, it must be located on the map by matching the window border color or numerical label. Next the operator must determine the orientation of the robot and its camera using cues such as prior direction of motion and matching landmarks between camera and map views. To gain this information the operator may choose to teleoperate the selected robot to locate it on the map, determine its orientation through observing the direction of movement, or simply to get a better viewing angle. The operator must then estimate the location on the map corresponding to the victim in the camera view. If "another" victim is marked nearby, the operator must decide whether the victim she is preparing to mark has already been recorded on the map.

Detecting and restoring a failed robot follow a similar time course: identifying the failed robot on the map and selecting it, then teleoperating it to its next waypoint where the automation can resume control.

The selected USAR environment was an office like hall with many rooms full of obstacles like chairs and desks. Victims were evenly distributed within the environment. Maps were rotated by 90º and each robot entered the environment from different locations on each trial. Because the laser map is built up slowly as the environment is explored and the office like environment provides few distinctive landmarks, there was little opportunity for participants to benefit from prior exposure to the environment. Robots followed predefined paths of waypoints, similar to paths generated by an autonomous path planner [7] to explore the map. All robots traveled paths of the same distance encountering the same number of victims and failures in each designed path. Upon reaching a failure point the operator needed to assume manual control to teleoperate the robot out of its predicament to its next waypoint where autonomous exploration resumed.

3 Experiment I

Experiment I reported in [8] compared a Control condition without alarms with two alarm conditions: Open-queue in which all malfunctions were displayed on a status panel and FIFO which displayed alarms one at a time in the order in which they occurred. Because all failures were of the same difficulty the order in which they are serviced should make no difference so according to the attention scheduling hypothesis the FIFO and open alarm conditions should produce equivalent performance. The experiment followed a three condition repeated measures design comparing the conventional MrCS displays with MrCS augmented by alarm panels. Conditions were fully counterbalanced for Map/starting points and display with 5 participants run in each of the six cells

3.1 Participants and Procedure

31 paid participants were recruited from the University of Pittsburgh community balanced among conditions for gender. None had prior experience with robot control although most were frequent computer users. Due to a system crash data was lost for one participant.

After providing demographic data and completing a perspective taking test, participants read standard instructions on how to control robots via MrCS. In the following 15 minute training session, participants practiced control operations. Participants were encouraged to find and mark at least one victim in the training environment under the guidance of the experimenter. After the training session, participants began the first 15 minute experimental session in which they performed the search task controlling 6 robots in the first assigned condition. At the conclusion of the session participants were asked to complete the NASA-TLX workload survey [9]. After brief breaks, the next two conditions were run accompanied by repeated workload surveys.

3.2 Results

Data were analyzed using a repeated measures ANOVA comparing search and rescue performance between the control and the two alarmed displays. No difference was found on the overall performance measures areas covered ($F1,29 = .488$, $p= .490$), victims found ($F1,29 = .294$, $p = .592$), or NASA- TLX workload survey ($F1,29 = 2.557$, $p = .121$). Significant effects were found on measures relating to operator strategy and the ways they performed their tasks.

Neglect time (NT) and latency in responding to failures are indicators of operator performance. Long NTs can indicate that some robots may have been ignored while latency in responding to failures can suggest noncompliance with assistance requests or heavy workload at other parts of the task. Robots in the FIFO condition were neglected longer than in the Control condition ($p = .033$, $SD = 619.507$) but did not differ significantly from the Open-queue condition. The neglect times were Open-queue= 1741, FIFO = 1887, and Control =1629 seconds.

Fault Detection time was defined as the interval between the initiating failure and the selection of the robot involved in that event. Cumulative Fault Detection times were significantly shorter for participants in the Alarm condition, p = .021, with a cumulative Fault Detection time of 933 seconds. Times for FIFO and Control conditions were 1120, and 1210 seconds respectively. A pairwise T-test shows a significant difference between the Alarm and Control conditions (p = .021, SD = 607.914).

Victim Delay time was defined as the interval between when a victim first appeared in a robot's camera and the selection of that robot. Victim Delay time again differed across conditions with average times of Open-queue 1303, FIFO 1548, and Control 1559 seconds. A pairwise T-test shows differences between Open-queue and FIFO (p= .041, SD =613.725), and Open-queue and Control conditions (p = .025, SD =578.945).

A related measure, Select-to-Mark, is defined by the interval between selecting a robot with a victim in view and marking that victim on the map by the process described earlier. Select to mark times can be interpreted as a measure of situation awareness (SA) because they require the operator to orient and interpret the environment. For this measure the results are reversed with users in the Open-queue condition taking the longest times (17.56 sec) and the Control the shortest (14.91 sec) with the FIFO condition (16 sec) again falling in between. There was no overall effect for select to mark time across the three experimental conditions (F(1.669,56) = 1.618, p = .212). A pairwise T-test, however, shows a difference between Open-queue and Control conditions (p = .025, SD = 6.02).

4 Experiment II

Experiment II reported in [10] extended the investigation begun in Experiment I by introducing multiple types of failures to allow a condition for which the schedule-aiding hypothesis would predict superior performance. Serving the shortest job first (SJF) is a provably optimal policy for maximizing throughput in a queuing system [5]. An alarm system that displayed only the current failure with the shortest time to repair, therefore, should improve the performance of the human-multirobot system over the Open-queue condition unless the unaided human is also following the same SJF policy.

4.1 Types of Failures

Recoverable failures were categorized into 4 major types, based on the data for commonly occurring non terminal and field repairable failures for the Pioneer P3-AT [11]. Two of these, camera and map failures, involve loss of display due to communication difficulties. The third, teleoperation lag is a control problem found by [12] to significantly degrade operator performance. The fourth, "stuck", is a common condition in which a robot becomes entangled with obstacles. To resolve encountered failures, the operator needed to manually guide the robot from its current location to

the next waypoint. Because each of the failure types imposed different difficulties for recovery, they took varying amounts of time to resolve. In order to estimate typical resolution times for different failures, a pretest using 10 participants was conducted as shown in Table 1 and Figure 2.

Table 1. Error Types

Failure	Description
Stuck	Robot was stopped by approaching obstacles
Teleoperation Lagged	Robot executed operator's command with 2~3 seconds delay
Camera Sensor Failed	Robot's video feed will be frozen right before the failure happened
Map Viewer Failed	Robot's position on the map viewer will be unable to update

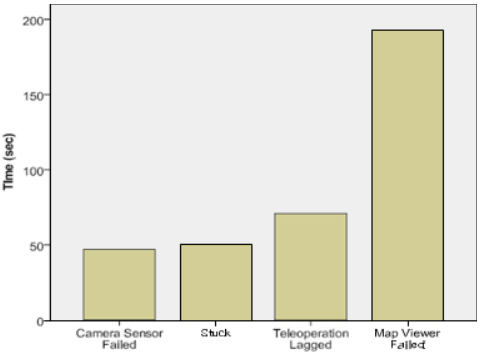


Fig. 2. Time to repair

In the training session, participants practiced control operations for different types of failures for 5 minutes each. Participants were instructed that their goal was to resolve failures by teleoperating to the next indicated waypoint as rapidly as possible. To avoid unrelated delays, such as those associated with switching attention among robots, participants controlled a single designated robot at a time. Because teleoperating the robot to its next waypoint was most easily accomplished by locating both on the map, loss of map indication proved to be the lengthiest failure to repair. The stuck condition which required extensive manual maneuvering and the camera failure that made obstacle avoidance more difficult were the easiest to overcome, with teleoperation delay falling in the middle. This ordering of estimated interaction times allowed failures to be presented to the operator in a priority queue following a shortest job first (SJF) discipline, known to maximize throughput [5].

4.2 Participants and Procedure

The experiment followed a three condition repeated measures design comparing the MrCS augmented by a status panel (Open-queue) with presentation of single alarms following either a FIFO or SJF policy. Thirty paid participants were recruited from the University of Pittsburgh community balanced among conditions for gender. None had prior experience with robot control although most were frequent computer users. Participants read standard instructions on how to control robots via MrCS. In the following 20 minute training session, 5 minutes for each type of failure, participants practiced control operations by resolving failures, three times for each type.

Participants were encouraged to find and mark at least one victim in the training environment under the guidance of the experimenter. After the training session, participants began the first 15 minute experimental session in which they performed the foraging task controlling 8 robots in their first assigned condition. Participants had been told the main task was to locate victims with detecting and resolving robot failures as a secondary task. At the conclusion of the session, participants were asked to complete the NASA-TLX workload survey [9]. After brief breaks, the next two conditions were run accompanied by repeated workload surveys.

4.3 Results

Victims Found & Distance Traveled. No difference was found for the number of victims identified ($F(2,58)=.110$, $p=.896$). Each victim marking was compared to ground truth to determine whether there was actually a victim near the location. If a mark was made further than 2 meters away from any victim or multiple marks for a single victim were found, the marks were counted as false positives. The number of false positives showed a main effect for queue condition ($F(2,58)=4.637$, $p=.014$). A pairwise T-test found a significant difference between Open-queue (1.13 false) and FIFO (2 false) conditions ($p=.030$), as well as a difference between SJF (1.2 false) and FIFO ($p=.012$). No differences were found between Open-queue and SJF.

Unmarked victims that had appeared within a robot's FOV (field of view) without being marked were counted as false negatives (misses). Operators in the Open-queue condition missed the most victims (15) and FIFO the fewest (11) with SJF falling in between (13). A repeated measures ANOVA shows a main effect among queue conditions, $F(2,58)=20.5$, $p<.001$. Pairwise T-tests revealed differences between Open-queue and FIFO ($p<.001$), Open-queue and SJF ($p=.006$), and SJF and FIFO ($p=.003$).

No difference was found for the distance traveled ($F(2,58)=1.73$, $p=.186$) although Open-queue (321m) appears slightly better than FIFO (293m) with SJF again in the middle (310m).

Neglect time (NT) and latency in responding to failures again served as indicators of operator performance. Long NTs can indicate that some robots may have been ignored while latency in responding to failures can suggest noncompliance with assistance requests or heavy workload at other parts of the task. NT ($F(2,58)=1.66$, $p=.20$) and the latency in responding to failure ($F(2,58)=1.75$, $p=.183$) were not significantly different among the three conditions. Pairwise T-tests found no difference between Open-queue and FIFO in either Neglect Time ($p=.086$) or fault detection time, ($p=.079$) although Experiment I had found longer NT in the FIFO condition.

The time to service failed robots, measured as the time between selecting the robot and resolving its problem again showed no difference among conditions ($F(2,58)=.579$, $p=.507$), which suggests the types of pre-designed failures were well distributed among three conditions. Overall, FIFO-queue appears slightly worse in the above three measurements.

Select-to-Mark, is defined by the interval between selecting a robot with a victim in view and marking that victim on the map. Select to mark times can be interpreted as a measure of situation awareness (SA) because they require the operator to orient and interpret the environment. A repeated measures ANOVA shows a significant difference among conditions ($F(2,58)=5.413$, $p=.011$). Operators in the

FIFO condition took the longest time (583 sec) and the Open-queue was the shortest (389 sec) with the SJF falling in between (478 sec). A pairwise T-test showed a significant difference between Open-queue and FIFO conditions ($p=.002$), and a marginal difference between Open-queue and SJF ($p=.061$).

The operator must successfully teleoperate the stopped robot from its current location to the next predefined waypoint to resolve a failure. A repeated measures ANOVA showed a significant difference for the count of resolved failures among experimental conditions ($F(2,58)=5.5$, $p=.006$). Participants in the Open-queue condition solved the most failures (17.8), which was significantly more than FIFO ($p=.003$). A pairwise T-test also revealed a difference between SJF, 17 failures, and FIFO 15.7 failures, ($p=.057$).

As in Experiment I the full-scale NASA-TLX workload measure found no advantage among conditions. To examine effects related to the highly prescriptive aiding in FIFO and SJF, we analyzed the frustration scale separately Repeated measures ANOVA showed a significant difference ($F(2,58)=5.159$, $p=.009$). Pairwise T- tests revealed differences between Open-queue and FIFO ($p=.038$) and between Open-queue and SJF ($p=.004$).

5 Discussion

In Experiment I we found that alerting operators to robots in need of interaction improved performance along a number of dimensions. The study compared a control condition without alerting with experimental conditions corresponding to the Open-queue and FIFO conditions of Experiment II. While alerting was beneficial, FIFO which directed the operator to service a particular robot was less effective than the Open-queue which allowed the operator to choose. This contradicts the predictions of the attention scheduling hypothesis which required that human attention be directed without loss of cognitive efficiency. The advantage for less constrained operators might be explained either by superiority of strategies of Open-queue operators when allowed choice or operator difficulties in complying with automation that prescribed the robot to be serviced.

Experiment II partially supported the premise that operator attention can be directed to interaction with individual robots without degrading performance. Open-queue performed slightly better than SJF on false positives, distance traveled, and failures resolved, but only for select-to-mark times did the difference approach significance. For the primary task of marking victims, FIFO participants proved slightly better, however, SJF participants were significantly superior to Open-queue users yielding a balanced performance which was never poorest. The above results may be due to the differences in allocation of attention. Within limited cognitive

capacity of processing information, operators have to selectively dedicate attention to any of the "wanted" targets and filter out the irrelevant information simultaneously [16]. Open-queue operators must devote time and attention to monitoring and selection of robots for servicing as well as the interaction leaving less available for the victim monitoring and marking tasks; whereas operators in the forced queue (Priority-/FIFO-Queue) conditions, by contrast, do not have to compete with monitoring and selecting robots to service leaving more resources available for victim-related tasks, which leads to the reversed results in unmarked victims among three conditions.

The FIFO-queue condition which directed operator attention suboptimally also led to the greatest loss of SA as reflected in its longest Select-to-Mark victims times and lowest marking accuracy. This may have been exacerbated by the FIFO discipline which did not distinguish between distracting recoveries such as loss of track on map and brief interventions such as maneuvering around an obstacle. For the Priority-queue, the SJF discipline had not only the advantage of allowing operators to work primarily on briefer interventions thereby preserving SA, but by clustering similar types of failures increased opportunities for reducing the cost to switch between recovery strategies and sharing the similar cognitive procedures among failures. However, the Priority-queue operators may have simply devoted more of their time and attention to robot requests than operators using the less efficient FIFO because of their greater payoff, which could be observed from the higher rate of unmarked victims.

Table 2 summarizes effects from the two studies. Performance on the primary victim detection and marking task was poorest in the Open-queue condition with more misses and fewer false alarms suggesting operators may have been devoting less effort to this task. When they did see a victim, however, they were faster to select the robot and mark the victim then those using priority queues indicating better situation awareness. This advantage extended to the secondary task where Open-queue users were faster to address and resolve faults. The performance improvements came at some cost, however, as indicated by the elevated frustration scale of the workload measure.

Taken together these experiments fail to confirm the attention scheduling hypothesis as the FIFO and SJF interfaces that dictated the malfunctioning robot to be serviced led to decreased cognitive efficiency as reflected in poorer performance in direct comparisons. Two possible explanations are that: 1) lack of volition in choice of robot to service led to inefficiencies due to task switching [13] and loss of situation awareness due to shifts of attention to potentially remote locations or 2) characteristics of the priority queue conflicted with operator's intentions leading to disuse [14] and hence poorer performance.

In forced queue conditions operators receive an explicit recommendation for the robot to assist. Under extreme stress or time pressured tasks, humans tend to defer to automation and rely on the system for making decisions [15]. This increased compliance under high workload could be especially beneficial to system performance where optimal strategies such as SJF can be used to steer operator attention. Although automated aids can reduce decisional load, they carry little additional information about other robots in need of assistance or the general state of the system. Operators therefore need to regain SA every time they switch to serve a new robot. While working from a forced queue, operators must match the alarmed robots to the thumbnails and/or maps, which could increase the cost in switching attention among failures and robots.

Table 2. Summary of Effects

	Experiment I	**Experiment II**	**Effects**
Primary Task Performance Measures			
Area Covered	No Effect	No Effect	No Effect
N of Victims Found	No Effect	No Effect	No Effect
False Positives	Not Tested	FIFO > (Open, SJF)	FIFO > (Open, SJF)
Misses	Not Tested	Open > SJF > FIFO	Open > SJF > FIFO
Victim Delay Time	Open < (Control, FIFO)	Not Tested	Open < (Control, FIFO)
Select to Mark Time	Open < Control	Open < (SJF, FIFO)	Open < (SJF, FIFO, Control)
Secondary Task Performance Measures			
Failures Resolved	Not Tested	(Open, SJF) > FIFO	(Open, SJF) > FIFO
Fault Detection Time	Open < Control		Open < Control
Full Task			
Neglect Time	FIFO > Control	No Effect	FIFO > Control
NASA-TLX Workload	No Effect	No Effect	No Effect
Frustration Subscale	Not Tested	Open > (FIFO, SJF)	Open > (FIFO, SJF)

The study results are promising for the prospects of improving HRI performance through scheduling operator attention. The improvement of performance in queuing discipline shows that forced queue aiding can be effectively used by operators and might even lead to superior performance under more complex conditions where choice among robot requests becomes more difficult.

Acknowledgments. This research has been sponsored in part by ONR Grant N0001409-10680.

References

1. Cao, Y.U., Fukunaga, A.S., Kahng, A.: Cooperative mobile robotics: antecedents and directions. Autonomous Robots 4(1), 7–27 (1997)
2. Crandall, J.W., Goodrich, M.A., Olsen, D.R., Nielsen, C.W.: Validating Human–Robot Interaction Schemes in Multitasking Environments. IEEE Transactions on Systems Man and Cybernetics Part A Systems and Humans 35(4), 438–449 (2005)
3. Scheutz, M., Kramer, J.: Reflection and Reasoning Mechanisms for Failure Detection and Recovery in a Distributed Robotic Architecture for Complex Robots. Components (1), 3699–3704 (2007)
4. Xu, Y., Dai, T., Sycara, K., Lewis, M.: Service Level Differentiation in Multi-robots Control. System, 2224–2230 (2010)

5. Garey, M.R., Johnson, D.S., Sethi, R.: The Complexity of Flowshop and Jobshop Scheduling. Mathematics of Operations Research 1(2), 117–129 (1976)
6. Carpin, S., Lewis, M., Wang, J., Balakirsky, S., Scrapper, C.: Bridging the gap between simulation and reality in urban search and rescue. In: Lakemeyer, G., Sklar, E., Sorrenti, D.G., Takahashi, T. (eds.) RoboCup 2006: Robot Soccer World Cup X. LNCS (LNAI), vol. 4434, pp. 1–12. Springer, Heidelberg (2007)
7. Chien, S.Y., Wang, H., Lewis, M.: Human vs. Algorithmic Path Planning for Search and Rescue by Robot Teams. Human Factors 54(4), 379–383 (2010)
8. Chien, S.-Y., Wang, H., Lewis, M., Mehrotra, S., Sycara, K.: Effects of Alarms on Control of Robot Teams. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 55(1), 434–438 (2011)
9. 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, vol. 1, pp. 139–183. North-Holland (1988)
10. Chien, S.Y., Mehrotra, S., Lewis, M., Sycara, K.: Scheduling Operator Attention for Multi-Robot Control. In: Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012), pp. 473–479 (2012)
11. Carlson, J., Murphy, R.R., Nelson, A.: Follow-up analysis of mobile robot failures. In: IEEE International Conference on Robotics and Automation, ICRA 2004, vol. 5, pp. 4987–4994 (2004)
12. Sheridan, T.B.: Space teleoperation through time delay: review and prognosis. IEEE Transactions on Robotics and Automation 9(5), 592–606 (1993)
13. Kiese., A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A., Koch, I.: Control and interference in task switching–a review. Psychological Bulletin 36(5), 840–874 (2010)
14. Kirlik, A.: Modeling strategic behavior in human-automation interaction: why an 'aid' can (and should) go unused. Human Factors 35(2), 221–242 (1993)
15. Inagaki, T.: Adaptive Automation: Sharing a Trading of control. In: Hollnagel, E. (ed.) Handbook of Cognitive Task Design (2003)
16. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans: A Publication of the IEEE Systems, Man, and Cybernetics Society 30(3), 286–297 (2000)

Explicit or Implicit Situation Awareness? Situation Awareness Measurements of Train Traffic Controllers in a Monitoring Mode

Julia C. Lo[1], Emdzad Sehic[1,2], and Sebastiaan A. Meijer[1,3]

[1] Faculty of Technology, Policy and Management, Delft University of Technology, The Netherlands
j.c.lo@tudelft.nl
[2] ProRail, The Netherlands
emdzad.sehic@prorail.nl
[3] Department of Transport Science, KTH Royal Institute of Technology, Sweden
smeijer@kth.se

Abstract. Railway traffic control faces the challenge of ensuring a high infrastructure capacity to maintain a constant train traffic flow. The current study assesses the situation awareness (SA), as a predictor of decision-making, of train traffic controllers to gain novel insights in their cognition. This study puts emphasis on levels of implicit and explicit situation awareness in a monitoring mode, through measures of SAGAT, MARS and performance. A human-in-the-loop simulator, called the PRL game is used to simulate the workspace of train traffic controllers. Initial findings indicate rather low levels of explicit SA, on the contrary to higher subjective SA scores through MARS and observer ratings, and a high performance on the punctuality and unplanned stops of trains.

Keywords: Situation Awareness, Implicit, Explicit, Train Traffic Control, Gaming Simulation.

1 Introduction

The railways in the Netherlands is characterized as the most dense and heavy utilized railway infrastructures in Europe [1]. The growing demand on diversified and higher frequency train schedules puts challenges on railway infrastructure innovations as well as on the implications of the work environment for train traffic and (regional and national) network controllers.

Railway infrastructure innovations are sought in processes as the increase of infrastructure capacity cannot be achieved by technical solutions alone [2]. Therefore, gaming simulations have been predominantly used so far as a research environment to test the viability of these process innovations. A gaming simulation can be seen as a simulation of a system, in which humans take part through game design methods and concepts such as immersion [3]. They comprise of different types, varying from

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 511–521, 2014.

human-in-the-loop alike simulators to table top games. For the Dutch railways, human-in-the-loop simulators are currently used for individual operators, whereas board or table top gaming simulations are applied to study a larger part of the railway traffic operating systems (for examples see [2]).

One of the key elements to determine the viability of the various innovation processes is the quality of decision-making of operators. Ultimately their decisions impact their performance and therefore might impact the performance of the railway traffic system. The cognitive concept of situation awareness is often addressed as a predictor of decision-making in complex socio-technical systems.

Situation awareness (SA) has been most widely defined as 1. the perception of elements in the environment, 2. the comprehension of these elements and 3. the projection of these elements in the near future [4]. Situation awareness can be seen as a dynamic mental model of the situation, in which explicit and implicit levels of knowledge can be distinguished [5]. Explicit knowledge refers to active knowledge that resides in the working memory, while implicit knowledge is refers to less active knowledge that cannot be inferred from queries or knowledge probes, being non-intentional, non-conscious and intuitive [6-8]. Following Croft, Banbury, Butler and Berry [6], implicit situation awareness can also be viewed as implicit processes in SA. Implicit processes are characterized as extremely durable, more robust in the light of competing attentional demands, and related to an increase in expertise. Measurements examples of implicit SA are described as comparisons of recalling probes (explicit SA) with performance-based measures or speed/accuracy measurements of elements (e.g. indicating hostile or friendly aircrafts) [6], [8-9].

The implications of cognitive concepts, such as situation awareness for decision-making, have been studied in a number of domains, e.g. aviation, nuclear power plants, military and firefighting [6]. However, previous studies identified different cognitive strategies across the domains. For example, air traffic controllers focus 90% of their time on processing information [10], while 95% of the tactical commanders in the military domain use a recognition decision strategy [11]. These different strategies can have implications on how the SA of operators is formed. Railway traffic controllers spend a significant amount of time in monitoring the train traffic flow. They operate in a fashion that requires only active involvement from the human operators when the delays of trains are affecting the train traffic flow, which might be caused by a failure in material, infrastructure or incidents. Findings from an ethnographical study on the decision-making of railway traffic controllers (i.e. train traffic and regional network controllers) reveal that operators do not always look in-depth into different consequences and have difficulties in making their reasoning explicit [12]. However, as far as known, no studies exist that investigate the SA of railway traffic controllers that make use of a traffic management system.

The findings from the ethnographic study raise the question how relevant cognitive constructs are formed, i.e. situation awareness, thus leading to the research question: *to what extent do train traffic controllers exhibit explicit or implicit levels of situation awareness*? The formulated question provides novel insights in the awareness levels of train traffic controllers and how this affects their performance in a monitoring mode of operations.

2 Method

2.1 Experimental Setting

As part of a suite of railway games, the human-in-the-loop simulator – PRL game – is used in this study. The inclusion of 'game' in its name, originated from previous design versions, which had abstract interfaces compared to the current version.

Aside from the current research focus, the general purpose of the study was formulated to investigate the impact of a gaming simulation session on the quality control processes of a new train timetable. Therefore, the session was also focused on the experience of the current human-in-the-loop simulator and to provide feedback regarding the new train timetable. Based on a set of gaming simulation components [2-3], Table 1 describes the characteristics of the gaming simulation NTTZ (new train timetable Zwolle).

Table 1. Characteristics of the PRL game NTTZ

Core aspect	*Description*
Purpose	Studying the impact of a game session on the quality control processes of a new train timetable
Scenarios	Two for each participant: 1. 2013 train timetable, 2. 2014 train timetable
Simulated world	Detailed infrastructure; detailed timetable; limited options in number of actions; larger area of train traffic operations (merged workspace of Zwolle station east-side and Zwolle station west-side)
# of participants	1 per session
Roles	Train traffic controller
Type of role	Similar to their own roles
Objectives	Execution of tasks – similar as to in their daily work
Constraints	Exclusion of roles outside the defined infrastructure area, exclusion of train driver, no large disruption
Load	Average train delays
Situation (external influencing factors)	Presence of individual observers seated next or near the participant, facilitators
Time model	Continuous

Both scenarios were designed together with subject matter experts to simulate a light disruption by mildly delayed trains, which was based on the realization data, i.e. average delays of trains in a month of that year with an average amount of disruptions. Further on, the first scenario focused on the 2013 timetable for participants to familiarize with the simulated environment and to obtain a base rating of SA and performance of train traffic controllers. Similarly, SA and performance were measured as well in the 2014 scenario.

As indicated in Table 1, the simulated workspace was represented by two merged workspaces surrounding the station of Zwolle with the borders of each workspace including smaller stations in their vicinity. In most of the cases, one train traffic controller is responsible for monitoring and controlling the train traffic flow at one workspace. Together, these workspaces form with two other workspaces (Hengelo and Deventer) the regional traffic control center in Zwolle.

Participants. Eleven train traffic controllers from one regional traffic control center in Zwolle took part in both scenarios. Train traffic controllers were selected based on their competence to operate at the simulated workspace.

2.2 Materials

A number of background questions were presented before each session: *work experience in the railway sector*, *work experience in the current job function*, *perceived experience of the workspace*, *perceived competences in comparison to peers, motivation in participation of the PRL game*. The latter three items were measured on a five-point Likert scale, varying from 'fully disagree/strongly unexperienced' to 'fully agree/strongly experienced'.

Multiple situation awareness measurement methods were used to triangulate measures of SA. Firstly, at the end and two times during each scenario, the gaming simulation was frozen, in which participants receive 22 queries in total, in accordance with the *Situation Awareness Global Assessment Technique (SAGAT).* In total three pauses were introduced, where seven to eight SAGAT questions were presented with a multiple-choice answering format (e.g. [13]). The pauses were planned after possible conflicting choices in the train traffic flow. The SAGAT queries were based on a goal-directed task analysis (GDTA) (e.g. [9]) from a national network controller. A subject matter expert translated the relevant SA requirements to a number of queries for a train traffic controller. Examples of the SAGAT questions are displayed in Table 2.

Table 2. Examples of the SAGAT queries

SA level	*Query example*	
1	At which track does train 13828 arrive in Zwolle?	Track 14, track 15, track 16
2	Which train leaves first according to planning from station Zwolle?	[Train number] 12522, 3629, 9119
3	How is the track capacity at 7:46 in station Zwolle?	6 tracks free, 5 tracks free, 4 tracks free, other, namely: ..

Four SA probes were removed during the analysis, which reduced the number of SAGAT queries to 19 for each scenario. For the analysis of the results, the items were firstly scored as correct/incorrect, and then the percentage of correct answers was calculated.

Subjective SA ratings were collected through the *Mission Awareness Rating Scale (MARS)* [3], [14] and presented at the end of each scenario. The selected MARS questions can be seen as the *perceived own situation awareness* and are related to the three SA levels as identified by Endsley [4], and respectively scored on a four-point scale, varying from 'fully disagree' to 'fully agree'. The average of the three levels of SA was calculated for the results.

Additionally, MARS-based questions were used to measure the *observed situation awareness*. One subject matter expert was present during all sessions, and rated the observed situation awareness with similar questions and a similar scale as the *perceived own situation awareness*. An observation sheet, which is also used by instructors during training session was provided as a guideline to rate the observed situation awareness of the participants. Similarly, the average of the three levels of SA was computed for the results.

In the railway sector, performance is measured on system level. That is, no official objective measurements for the individual performance of train traffic controllers exist. The performance indicators 'punctuality' and 'unplanned stops' were identified in consultation with the railway performance & analytics department. These results were retrieved from log files of the PRL game. Punctuality can be defined by the entry and exit times of trains within a specific region for a specific workspace. The punctuality of trains is often measured in percentages over a certain amount of time and for a certain level of delay. As the performance indicator for the Dutch railway infrastructure organization is set at three minutes, this is also used in the current setup. Secondly, unplanned stops are defined as the number of times that train drivers encounter an unplanned red signal. Unplanned stops can be seen as hazardous for safety, therefore the reduction of unexpected red signals is wished to be achieved.

The human-in-the-loop experience of train traffic controllers was captured by measurements of *gaming simulation validity* [15] at the end of each session. In line with Raser [15], three out of four components of gaming simulation validity were included, namely structural validity (similarity in structure between the simulated and reference systems), processes validity (similarity in processes between the simulated and reference system), and psychological reality (the degree to which the participants/players perceive the simulated system as realistic). Three items measured structural validity (α = .60). An example of an item was: 'I can apply the information from the information sources in the simulator in a similar way as in the real world'. Process validity was measured as well by three items (α = .90), e.g. 'the train traffic flow in the simulator is similar in their processes to the real world train traffic flow'. Thirdly, psychological reality was measured by seven items (α = .84), e.g. 'the train model is sufficiently realistic for the current task'. All items were measured on a five-point Likert scale, varying from 'fully disagree' to 'fully agree'.

Five *workload* items from the NASA-TLX [16] were presented after each scenario (α = .64). One item related to the physical demand was removed as this was unnecessary to measure given the task at hand. In line with the scales used at other items, a five-point Likert scale was applied.

Lastly, the difficulty between the two scenarios was measured on a similar five-point Likert scale in the post-session questionnaire, in which the statement was provided: 'I was able to quickly get accustomed to the new timetable'.

2.3 Procedure

The sessions took place over three days. All train traffic controllers conducted two scenarios: one scenario with the current 2013 train timetable and one scenario with the new 2014 train timetable, in which the length of each scenario was 35 minutes. At the start of the session, participants firstly received instructions on the possibilities and limitations of the PRL game. Additionally, they were also asked for permission to record the session by video. Subsequently, participants received the pre-questionnaire. 15 minutes after the start of the scenario, the PRL game was paused, in which participants were asked to turn to the desk behind them and answer the SA probes. Two more freezes of the PRL game followed every 10 minutes, in which the third pause marked the end of the scenario. Similar halts in the scenarios were introduced in the second scenario, which ended with a post-questionnaire. In both scenarios, observers asked questions and feedback about the usability of the PRL game, their decision-making and their preferences related to details in the timetables, during segments of the scenarios where train traffic controller was monitoring the train traffic flow. Conversations during the operator's task in a non-severe disruption were planned in the procedure as they are seen as consistent with the displayed behavior of train traffic controllers in their work environment.

Fig. 1. Two PRL game sets: in the foreground the PRL game set with the new timetable, on the background the set with the current timetable

3 Results

Ten male train traffic controllers and one female operator took part in the sessions. Their work experience within the current job function was 16.5 years (*SD* = 8.9), however the average of their overall work experience in the railway sector was higher, *M* = 20.8, *SD* = 9.8. Participants indicated that they perceived their competence on the current workspace as high (*M* = 4.3, *SD* = .65), as well in comparison to colleagues within the regional traffic control center (*M* = 3.8, *SD* = .75). A high interest was indicated to participate in the PRL game (*M* = 4.4, *SD* = .92).

3.1 Perceived Difficulty of the Scenarios and Workload

As expected, the second scenario was not perceived as challenging to the train traffic operators, as they were able to quickly get accustomed to the new timetable (*M* = 4.2, *SD* = 2.1). Additionally, participants indicated that they perceived a low workload *(M* = 1.7, *SD* =.47). Qualitative data obtained during the session supported both results.

It should be remarked that the scenario was initially designed to simulate a light disruption. However, train traffic controllers perceived the delays of the trains as not sufficiently problematic to interfere with the train traffic flow, possibly due to the fact that the automation function of the train traffic system ('ARI') was not triggered by the current train delays to indicate red highlights on the timetable screen (i.e. ARI is not able to manage to execute the train path assignment of a train). Therefore, there was less interaction between the operators and the PRL game than expected, making it a monitoring mode instead.

3.2 Gaming Simulation Validity

The PRL game is validated for the use of the current task through the gaming simulation validity dimensions, structural validity, process validity and psychological reality. The results indicate a rather positive perception of the PRL game by the train traffic controllers (see Table 3). Qualitative data support this notion as well. The rather positive scores on the three validity types also assert the assumption that levels of situation awareness and performance in the simulated environment should be comparable to a similar task in their real work environment.

Table 3. Gaming simulation validity dimensions of the PRL game for the current task

	N	***M***	***SD***
Structural validity	11	3.6	.53
Process validity	11	3.7	.71
Psychological reality	11	3.8	.53

3.3 Situation Awareness and Performance

Results from measurements on situation awareness and performance are depicted in Table 4. The lower number of participants in the scenarios can be subscribed to missing, unclear or discarded data (e.g. by deviation from the probes instructions).

Table 4. Measurements of situation awareness and performance

	Scenario 1			*Scenario 2*		
	N	**M**	**SD**	**N**	**M**	**SD**
SAGAT (%)	9	44.4	17.68	9	37.11	11.07
Perceived SA (1-4)	8	3.1	.59	11	3.3	.49
Observed SA (1-4)	11	3.5	.43	10	3.7	.41
Punctuality (%)	11	99.3	1.20	11	98.6	1.64
Unplanned stops	11	2.2	.87	11	2.4	1.29

As the variables punctuality, perceived situation awareness, observed situation awareness (second scenario) and unplanned stops (first scenario) were significant on the Kolmogorov-Smirnov statistic for normality, a Wilcoxon test was conducted to analyze differences between scenarios in scores. A significant difference was found for the observed SA scores ($Z = -2.33$, $p = .02$).

The values in the table indicate rather low explicit SA levels measured through SAGAT. However, subjective ratings by participants themselves and observers show high to very high levels of situation awareness. Also for punctuality, the results indicate near optimal achievements by the train traffic controllers.

Correlations were drawn between the variables to investigate the relation between situation awareness and performance. Relevant trends were found for a relation between explicit situation awareness and punctuality in the first scenario ($\rho = .64$, $p = .06$); a higher explicit situation awareness leads to a higher punctuality of trains. Additionally, trends were found for the perceived situation awareness and both performance indicators punctuality and unplanned stops in the second scenario (respectively ($\rho = .53$, $p = .09$; $\rho = -.52$, $p = .10$). A higher perceived SA leads to a higher punctuality, and a higher perceived SA leads to less unplanned stops of trains.

The findings reveal that train traffic controllers perceived their SA as high and show a high performance in the current monitoring mode. However, the train traffic controllers score reasonably low on objective (explicit) measures of SA, which is in line with the implication for the presence of implicit situation awareness. This finding might be supported by the results on the SA probes that are categorized by SA level (see Table 5).

The results show that SA level 1 items were fairly low and beneath level 2 scores, although absolute values of the level 2 items remained rather low. Theoretically, SA would drop with each SA level, i.e. operators set a baseline for their SA by the elements they perceive in the situation. Following this, their comprehension (SA level 2) is equal to lower than their perception, and similarly their projection (SA level 3) is equal or lower than their comprehension.

Table 5. SA probes per SA level

	Scenario 1		*Scenario 2*	
	# total items (N=9)	**% correct**	**# total items (N=9)**	**% correct**
SA level 1	99	37	108	39
SA level 2	54	65	45	42
SA level 3	18	39	18	17

Additionally, implicit SA could also be measured by using explicit SA measures and relating this to certain conflicting choices in the scenario. However, since the operators did not show much interference with the expected conflicts in the scenario and the train traffic flow in general, no implications could be drawn with regards to their SA queries and behavior in the scenarios.

Explorative analyses were conducted to investigate the potential role of individual differences between train traffic controllers. A trend was found for a negative correlation between the work experience in the railway domain and percentage of correct SAGAT answers; $r = -.65$, $p = .06$; a higher experience in the railway domain lead to a lower explicit situation awareness.

4 Discussion and Conclusion

The current study attempts to investigate the level of explicit and implicit situation awareness at train traffic controllers. The findings show rather low levels of explicit SA, on the contrary to (very) high subjective SA scores through MARS and observer ratings, and high performance on the punctuality and unplanned stops of trains. It is possible that the low explicit SA scores are influenced by the fact that the light disruption was not seen as that problematic and operators therefore portrayed monitoring modes of operation. This changed mode of operation might also have been affected by the role of automation as train traffic controllers rather would rely on and are triggered by the automation function of the train traffic system, causing a low explicit SA. Similar findings with low SAGAT scores have been found in earlier studies where no active decision-making was taking place [17]. Another possible explanation for the low explicit SA might be related to relevance of the presented SA queries in the unexpected changed mode of operation. It is possible that certain information is less relevant as different goals are achieved in certain circumstances, e.g. operators search for deviations and irregularities in the train traffic flow.

Further on, a trend is found for the negative relation between work experience and explicit SA. This result might be in line with the notion from previous studies that an increased implicit SA is more common for expert operators, e.g. [6]. Additionally, trends were found for the relation between objective (explicit) ratings of situation awareness and performance, but not found for both scenarios however, limiting the generalizability of these results.

Another limitation of the study is that no official performance indicators could be used, as these do not exist within the railway infrastructure organization. As critical remarks can be drawn for both performance indicators, more investigation is needed

to determine the theoretical and computational implications of individual performance indicators in the railway domain.

Nonetheless, these findings reveal novel descriptions on the situation awareness of train traffic operators in a monitoring mode. Further research is needed to identify the levels of explicit situation awareness in high disrupted conditions and to explore its relationship with levels of expertise.

Acknowledgements. This research was funded through the Railway Gaming Suite program, a joint project by ProRail and Delft University of Technology.

References

1. Ramaekers, P., De Wit, T., Pouwels, M.: Hoe druk is het nu werkelijk op het Nederlandse spoor? Het Nederlandse spoorgebruik in vergelijking met de rest van de EU-27. [How busy is it really on the Dutch railway tracks?]. Centraal Bureau voor de Statistiek (2009)
2. Meijer, S.A.: Introducing gaming simulation in the Dutch railways. Procedia - Social and Behavioral Sciences 48, 41–51 (2012)
3. Lo, J.C., Meijer, S.A.: Measuring group situation awareness in a multiactor gaming simulation: A pilot study of railway and passenger traffic operators. In: Proceedings of the 57th Human Factors and Ergonomics Society Annual Meeting, pp. 177–181 (2013)
4. Endsley, M.R.: Design and evaluation for situation awareness enhancement. In: Proceedings of the 32nd Human Factors Society Annual Meeting, pp. 97–101. SAGE Publications, Boston (1988)
5. Adams, M.J., Tenney, Y.J., Pew, R.W.: Situation awareness and the cognitive management of complex systems. Human Factors 37, 85–104 (1995)
6. Croft, D.G., Banbury, S.P., Butler, L.T., Berry, D.C.: The role of awareness in situation awareness. In: Banbury, S., Tremblay, S. (eds.) A Cognitive Approach to Situation Awareness: Theory and Application, pp. 82–103. MPG Books Ltd., Bodmin (2004)
7. Endsley, M.R.: The role of situation awareness in naturalistic decision making. In: Zsambok, C.E., Klein, G. (eds.) Naturalistic Decision Making, pp. 269–283. Lawrence Erlbaum Associates, Mahwah (1997)
8. Gugerty, L.J.: Situation awareness during driving: Explicit and implicit knowledge in dynamic spatial memory. Journal of Experimental Psychology: Applied 3, 42–66 (1997)
9. Endsley, M.R.: Direct measurement of situation awareness: Validity and use of SAGAT. In: Endsley, M.R., Garland, D.J. (eds.) Situation Awareness Analysis and Measurement, pp. 147–174. LEA, Mahwah (2000)
10. Kaempf, G.L., Orsanu, J.: Current and future applications of naturalistic decision making in aviation. In: Zsambok, C.E., Klein, G. (eds.) Naturalistic Decision Making, pp. 81–90. Lawrence Erlbaum Associates, Mahwah (1997)
11. Kaempf, G.L., Klein, G., Thordsen, M.L., Wolf, S.: Decision making in complex naval command-and-control environments. Human Factors 38, 220–231 (1996)
12. Steenhuisen, B.: Competing Public Values: Coping Strategies in Heavily Regulated Utility Industries. Gildeprint Drukkerijen, Enschede (2009)
13. Strater, L.D., Endsley, M.R., Pleban, R.J., Matthews, M.D.: Measures of platoon leader situation awareness in virtual decision-making exercises. U.S. Army Research Institute for the Behavioral and Social Sciences (2001)

14. Matthews, M.D., Beal, S.A.: Assessing situation awareness in field training exercises, U.S. Army Research Institute for the Behavioral and Social Sciences (2002)
15. Raser, J.C.: Simulations and Society: An Exploration of Scientific Gaming. Allyn & Bacon, Boston (1969)
16. 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, Amsterdam (1988)
17. Endsley, M.R., Rodgers, M.D.: Distribution of attention, situation awareness, and workload in a passive air traffic control task: Implications for operational errors and automation (No. DOT/FAA/AM-97/13). Federal Aviation Administration Office of Aviation Medicine (1997)

Two Types of Cell Phone Conversation Have Differential Effect on Driving

Weina Qu, Huiting Zhang, Feng Du, and Kan Zhang

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

Abstract. It has been demonstrated that cell-phone conversations impair driving performance. However, it is unclear whether the difficulty of naturalistic phone conversations can modulate driving performance. The present study used a simulator to investigate whether the cognitive load of phone conversations (non-conversation, multiple choice and open question conversations) influence highway driving performance. The results showed cell phone conversations with open questions led to most aggressive driving with highest speeds and shallowest braking. Furthermore, open question conversations led to the smallest route deviations. These results suggested that a drivers' capability for monitoring speed and making manoeuvre decreases as the difficulty of a phone conversation increases. The implications of this study for driving safety are discussed.

Keywords: Driving, Cell phone, Naturalistic conversations, Simulator, Open questions, Multiple choices.

1 Introduction

It is widely accepted that cell phone conversation has a negative impact on driving performance in terms of slower reaction time to immediate events [1, 3, 4, 8], higher variation in accelerator pedal position, slower and more variable speed, and higher level of workload [16]. Cell phone conversations also exposes drivers to higher risk of accidents [20]. Based on cell phone billing records, studies found a fourfold increase in the risk of car crash associated with phone use irrespective of the age and gender of the drivers [13].

Cell phone interference is mainly due to the cognitive load imposed by generating language rather than listening comprehension or repeating words. For example, a study showed that the costs of cell-phone conversations are not due to dual-task costs associated with listening to verbal contents or speaking words [21]. In addition, many studies also showed that cell phone interference is not due to manual operation of phones. Phone conversations, regardless of phone type (handheld and hands-free), have negative impacts on performance especially in detecting brief events [8, 9]. Handheld and hands-free phones produced similar RT decrements [3]. Regardless of the causes of cell phone interference on driving, a study showed that the driving

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 522–532, 2014.

impairment associated with cell phone conversations was comparable to that of drunken driving [20].

Other cognitive tasks, such as word tasks [21] and mathematics games [14], also interfere with driving. Studies have found that driving performance generally turns worse as the cognitive tasks become more difficult [2, 14, 17]. However, it is still unknown whether these results can be generalized to the effect of cell-phone conversations on driving.

Several studies examined whether the difficulty of cell phone conversations can modulate the driving performance [5, 11, 16, 19]. For example, a study manipulated the difficulty of conversations by using two sets of conversational topics rated by pilot testing as easy and difficult conversations, but failed to find any difference in driving performance between the easy and difficult conversations [16]. Another study assigned straightforward topics such as personal background and hobbies to the "simple" conversation condition, and topics that required more thought, such as mathematical calculation or logical reasoning, were considered as the ''complex'' conversations condition [11]. Again, Liu (2003) did not find differences in driving performance between the ''simple'' and "complex'' conversation conditions. It is still unknown why the difficulty of cell phone conversations does not modulate driving performance. A possible explanation is that complex conversations do not necessarily make participants engage in more demanding cognitive processing. Another possibility is that the content of different topics in conversations varied dramatically, so that participants' familiarity with those topics varied and their willingness to talk about various topics was also affected. These confounding variables might mask the effect of the difficulty of conversations.

What might make the situation worse is that people are overconfident with their driving performance. They are not fully aware of the potential risk associated with cell phone conversations while driving. A study showed that many drivers may not be aware of their decreased driving performance when using cell-phones [10]. Female drivers especially tended to have a greater discrepancy between the perception of their own driving performance and their actual performance.

The current study was designed to examine whether the difficulty of conversations might influence driving performance. A review paper indicated the use of driving simulator provides the most effective and most ethical method of the influence of mobile phone use on driving performance [7]. So in this study, two sets of naturalistic conversation questions, which were similar in length and addressing the exactly same set of topics, were presented via a hand-free cell phone to drivers who were driving on a high fidelity diving simulator. One set was comprised of multiple choice questions with two options, and the other set was comprised of open questions concerning the same set of topics as those in the multiple choice questions. Although the content was the same for both sets of questions, the open questions provided wider and more in depth information and required more cognitive processing than do the multiple choice questions [6, 15]. Participants had to come up options first for the open questions in relative to the multiple choices questions. Thus we expected the open questions to impose a higher cognitive load on drivers and impair their driving performance relative to multiple choice questions.

2 Method

2.1 Participants

12 participants (6 males and 6 females) took part in the experiment for monetary compensation. Their mean age was 33.4 years old, ranging from 23 to 55 years. All participants had valid driver licenses and at least 1 year of driving experience (their mean driving history was 77.7 months). Subjects reported having normal or corrected-to-normal vision. All participants owned a cell phone, and 10 participants reported that they have used a cell phone while driving.

2.2 Instruments

Driving Simulator
An interactive cockpit driving simulator (Sim-Trainer, designed by Beijing Sunheart Inc.) was used in the present study. The simulator has a 120° view, side view mirror, dashboards and manual transmission. Participants can drive by using the steering wheel, brake and accelerator pedals.

Driving Environment
The simulated highway driving environment for the experiment was a 3-lane highway. There was a 2km-long obstruction on the highway, so that drivers had to switch lanes there. Other than those switches, participants were told to maintain their lane and speed at 50km/h all the time. There was no other traffic in order to create a simple driving environment.

Conversation Questions
There were two different types of conversation question (multiple choice and open questions), which shared the same set of topics. The multiple choice questions contained two options for the participants to choose from, and the choice was set to be obvious and easy to make, based on the results of the pilot study.

The order of the types of conversation question was counterbalanced between participants so that half of the participants answered all the multiple choice questions first, and the other half answered all the open questions first. The participants were first asked to answer multiple choice or open choice questions, then they were asked to explain why they made their choice with no more than five sentences; otherwise, the experimenter would interrupt them and move on to the next question.

Additionally, the survey collected some of participants' demographic information considered to be related to driving safety including gender, age, driving years, overall mileage, and weekly mileage.

2.3 Design

There were three conditions: non-conversation, multiple choice questions and open questions. Participants performed the same driving task for all three conditions. They

first drove under the non-conversation condition without any cell phone conversation. Then they were required to drive the same route under the multiple choice questions and open questions conditions. These two conditions were counterbalanced across participants. In addition, all participants were asked to maintain the speed at 50 km/h and to stay in the lane they started with. Each condition lasted approximately 7 minutes. Since each participant drive the same route three times within half an hour, a possible practice effect is of major concern. To ameliorate this issue, participate were always required to start with the non-conversation condition. If a practice effect played an important role in driving performance, participants should have the better driving performance in two dual-task conditions (multiple choice questions and open questions) compared with the non-conversation condition. However, if driving performance in dual-task conditions is worse than the non-conversation condition, the impact of secondary task on driving might be stronger than observed results because practice effect might partially offset distracting effect of secondary task.

2.4 Procedure

Participants first drove a practice session to get used to the driving simulator and route. Participants used a Nokia C5110 cell phone with a blue-tooth headset. Participants were asked to maintain their speed at 50km/h. After the practice session, participants drove the same route with no conversation. Then they performed two more rounds of driving, each of which was assigned to either the multiple choice questions condition or open questions condition. The two conditions were counterbalanced across participants. The experimenter asked 13 questions in each condition via cell phone. Finally participants completed the subjective evaluation questionnaire.

2.5 Model for Analysis

The driving performance was evaluated by the speed maintenance performance (average speed, speed variability, average position of accelerator, accelerator position variability, average position of brake and brake position variability) and the lane keeping performance (average offset of steering wheel, variability of the offset of the steering wheel, average lateral deviation and lateral deviation variability). Repeated measures ANOVAs were conducted to test whether different types of conversation affected the speed maintenance and lane keeping.

3 Results

3.1 Speed Maintenance (Mobility)

A separate contrast analysis was conducted for each of the hypothesized effects for the speed maintenance variables.

Speed

As shown in Figure 1, types of cell phone conversation significantly influenced mean speed ($F(2, 22) = 8.065$, $p = .002$), with the fastest speed in the open question condition and the slowest speed in the non-conversation condition. Multiple comparison revealed that mean speed in the non-conversation condition was significantly slower than those in the multiple choice condition ($p = .070$) and the open questions condition ($p = .005$). The mean speed in the multiple choice questions condition was also slower than that of the open questions condition ($p = .009$).

The speed variability is shown in Figure 1b. Types of cell phone conversation significantly influenced the variability of speed ($F(2, 22) = 4.585$, $p = .022$). Multiple comparison revealed a significant difference between the non-conversation condition and the open questions condition ($p = .008$), but there was no other significant effect ($p > .100$).

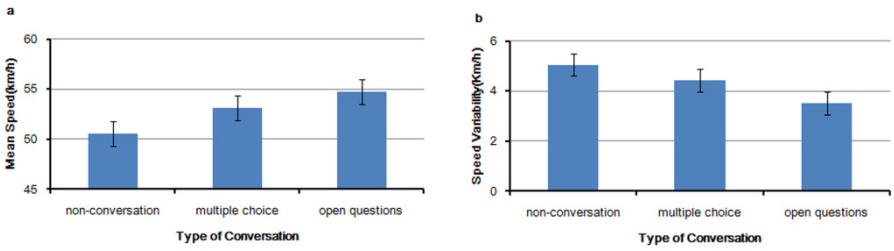


Fig. 1. Mean speed and speed variability of the three conditions. (a) Mean speed of the three conditions. (b) Speed variability of the three conditions. The error bars represent standard error.

Accelerator Position

Type of cell phone conversation did not have an effect on the mean accelerator position ($F(2, 22) = 1.981$, $p = .162$) or the variability of accelerator position ($F(2, 22) = 2.707$, $p = .089$).

Brake Position

The mean brake position is shown in Figure 2a. There is a main effect of the type of cell phone conversation on the mean brake position ($F(2, 22) = 4.946$, $p = .017$), with the shallowest brake position in the open question condition and the heaviest brake position in the non-conversation condition. Multiple comparison revealed a significant difference between the non-conversation and the open questions ($p = .007$), but no other significant effects ($p > .140$).

The mean brake position is shown in Figure 2b. There is a main effect of the type of cell phone conversation on brake position variability ($F(2, 22) = 4.454$, $p = .024$). Multiple comparison revealed a significant difference between non-conversation and the open questions ($p = .020$), and a marginally significant difference between non-conversation and the multiple choice questions ($p = .073$), but there was no significant difference between the multiple choice questions and the open questions ($p = .231$).

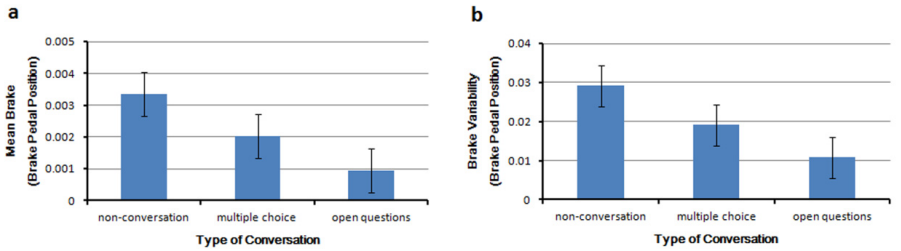


Fig. 2. Mean brake position and brake position variability of the three conditions. (a) Mean brake position of the three conditions. (b) Brake position variability of the three conditions. The error bars represent standard error.Lane position maintenance (Stability).

Deviation

The absolute deviation from route is shown as a function of conversation type in Figure 3a. The type of cell phone conversation significantly influenced the mean of the absolute deviation ($F(2, 22) = 20.971$, $p = .000$), with the smallest deviation in the open question condition and the largest deviation in the non-conversation condition. Multiple comparison revealed a significant difference between the non-conversation condition and the open question condition ($p = .000$). There was also a significant main effect on mean deviation between the non-conversation condition and the multiple choice questions condition ($p=0.009$). These results indicate that participants were more likely to deviate from the route in the non-conversation condition than in the open questions or multiple choice question condition.

The variability of absolute deviation from route is shown as a function of conversation type in Figure 3b. There was a main effect of cell phone conversation on the variability of absolute deviation ($F(2, 22) = 3.985$, $p = .033$). Multiple comparison revealed a significant difference between the non-conversation condition and the open questions condition ($p = .025$), but there were no other significant effects ($p > .100$).

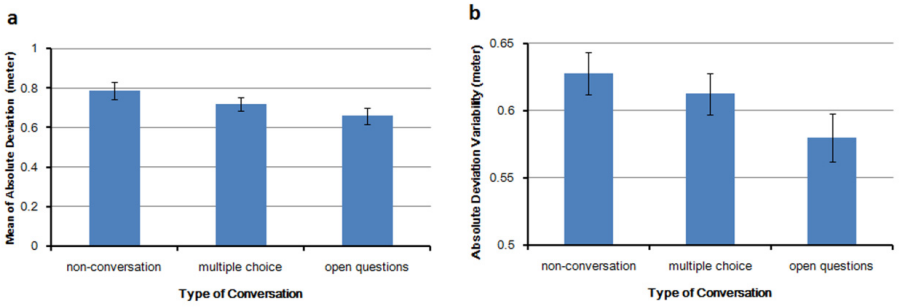


Fig. 3. Mean absolute deviation and absolute deviation variability of the three conditions. (a) Mean absolute deviation of the three conditions. (b) Absolute deviation variability of the three conditions. The error bars represent standard error.

Steering Offset

The offset of the steering wheel is shown as a function of conversation type in Figure 4a. Type of cell phone conversation significantly influenced mean steering offset ($F(2, 22) = 8.216$, $p = .002$), with the smallest steering offset in the open question condition. Multiple comparison revealed a significant difference between the non-conversation and the open questions conditions ($p = .023$) and between the multiple choice questions and the open questions ($p = .001$), but there was no significant effect between the non-conversation and the multiple choice questions conditions ($p = .631$). The variability of steering offset is also shown as a function of conversation type in Figure 4b. There was a main effect of phone conversation type on the steering offset variability ($F(2, 22) = 6.374$, $p < .001$). Multiple comparison revealed a significant difference between the non-conversation and the open questions conditions ($p = .027$) and between the multiple choice questions and the open questions ($p = .004$), but there was no significant effect between the non-conversation and the multiple choice questions conditions ($p = .510$). These results indicate that participants were least likely to make manoeuvres in the open questions condition.

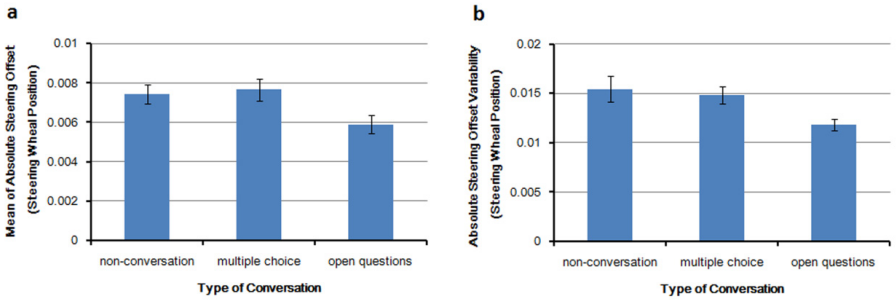


Fig. 4. Mean absolute steering offset and absolute steering offset variability of the three conditions. (a) Mean absolute steering offset of the three conditions. (b) Absolute steering offset variability of the three conditions. The error bars represent standard error.

3.2 Subjective Evaluation

Twenty out of twenty-four pilot participants (83.3%) rated the open questions more difficult than the multiple choice questions. Other three participants reported that two kinds of conversation questions were equally difficult. Only one participant rated the open questions easier than the multiple choice questions. These difficulty ratings were made when the pilot participants were not driving, and thus focused exclusively on the conversation task.

Participants evaluated the distracting effect of phone conversations on a seven-point scale ranging from "absolutely no disturbance" to "serious disturbance". The distracting effects of the two kinds of questions are listed in Figure 5. Participants rated the open questions condition (mean ± SE: 4.92 ± .50) much more distracting than the multiple choice questions condition (mean ± SE: 3.58 ± .45), and the difference was significant ($t(11) = 2.402$, $p = .035$).

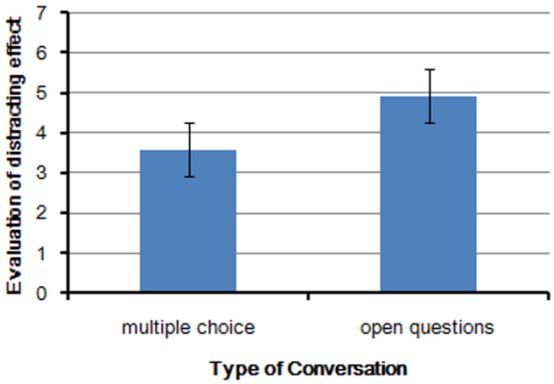


Fig. 5. The subjective evaluation of distracting effect of the conversation questions

Participants were required to evaluate their own driving performance for each of the three conversation conditions (non-conversation, multiple choice and open questions conversation) on a seven-point scale ranging from "very bad" to "very good". Their evaluation of driving performance is illustrated as a function of conversation type in Figure 6. The driving performance in the non-conversation condition was reported best (mean ± SE: 5.92 ± .34), followed by the driving performances in the multiple choice questions condition (mean ± SE: 4.50 ± .36). The participants reported that the driving performance in the open questions condition was the worst (mean ± SE: 3.75 ± .46). These subjective evaluations were submitted to a repeated measure ANOVA. There was a significant difference between the three conversation conditions, $F(2, 22) = 9.702$, $p = .001$, indicating that participants were aware of the strongest distracting effect in the open questions conversation condition.

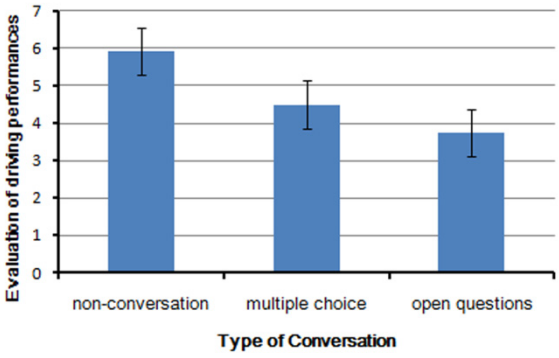


Fig. 6. The subjective evaluation of driving performance

4 Discussion

The present study replicated the classic finding of cell-phone conversation's interference with driving performance. More specifically, participants lose precise control of speed and tend to speed up when they are engaged by phone conversations. More importantly, this study clearly showed that participants tend to drive faster in the open question condition than in the multiple choice condition. In other words, phone conversations cause larger impairment in driving performance as the conversation becomes harder. This finding indicates that the difficulty of a cell-phone conversation modulates the cognitive load of the conversation, thus resulting in differential impairment on driving.

Presumably cell-phone conversation interferes with driving because the conversation engages cognitive processes which are attention consuming [18, 21]. Previous studies showed that some processes involved with cell-phone conversation, such as manual operation of cell phones, listening to verbal stimuli and repeating words, do not interfere with driving [3, 8, 9, 21]. The present results indicate that generating response options might be one of those attention-demanding processes that can interfere with driving. The present study is the first one to identify a specific cognitive process in cell-phone conversation that can interfere with driving.

The present study found that drivers speeded up when they are actively engaged in cell phone conversation. This finding was inconsistent with some previous findings that drivers slow down when they are talking on cell phones [16]. The discrepancy might be due to the task requirement of maintaining 50km/h on a traffic free highway in the present study. Since there is no traffic on the highway, participants might feel compelled to drive fast to satisfy the task demand. Thus participants depress the accelerator to almost the same extent across the three conversation conditions, but participants used the brake progressively less often as the phone conversation became harder. As a result, they were unable to monitor and control speed precisely as their attention was increasingly engaged in the cell phone conversation. In other words, the heavier load was imposed by the cell phone conversation, the faster they drove.

Participants also tended to more strictly stick to their route and to maneuver less as the conversation became harder. This further demonstrated that, when drivers are progressively engaged by conversation, they lose their ability to closely monitor their driving status. This impairment may increase their risk of having a traffic accident in more complex situations.

Drivers in the present study were able to perceive the disruptive effect of cell phone conversation on driving. However, most of them except one still reported cell phone usage in driving. Actually it is common among drivers who often feel compelled to take a phone call during driving [22]. This study indicated that drivers should try their best to avoid very complex cell phone conversations involving generating multiple options.

Acknowledgements. This study was supported by grants from the National Natural Science Foundation of China (Grant No. 31200766 and No. 91124003) and the Scientific Foundation of the Institute of Psychology, Chinese Academy of Sciences (Grant No. Y1CX212005). We thank Beijing Sunheart Simulation Technology Ltd. for their technical support. We would also like to thank Richard Carciofo for proof reading.

References

1. Beede, K.E., Kass, S.J.: Engrossed in conversation: The impact of cell phones on simulated driving performance. Accident Analysis and Prevention 38, 415–421 (2006)
2. Briem, V., Hedman, L.R.: Behavioural effects of mobile telephone use during simulated driving. Ergonomics 38, 2536–2562 (1995)
3. Caird, J.K., Willness, C.R., Steel, P., Scialfa, C.: A meta-analysis of the effects of cell phones on driver performance. Accident Analysis and Prevention 40, 1282–1293 (2008)
4. Collet, C., Guillot, A., Petit, C.: Phoning while driving I: a review of epidemiological, psychological, behavioural and physiological studies. Ergonomics 53(5), 589–601 (2010a)
5. Collet, C., Guillot, A., Petit, C.: Phoning while driving II: a review of driving conditions influence. Ergonomics 53(5), 602–616 (2010b)
6. Friborg, O., Rosenvinge, J.H.: A comparison of open-ended and closed questions in the prediction of mental health. Quality & Quantity 47(3), 1397–1411 (2013), doi:10.1007/s11135-011-9597-8
7. Haigney, D., Westerman, S.J.: Mobile (cellular) phone use and driving: a critical review of research methodology. Ergonomics 44(2), 132–143 (2001)
8. Horrey, W.J., Wickens, C.D.: Examining the Impact of Cell Phone Conversations on Driving Using Meta-Analytic Techniques. Human Factors 48(1), 196–205 (2006)
9. Ishigami, Y., Klein, Y.M.: Is a hands-free phone safer than a handheld phone? Journal of Safety Research 40, 157–164 (2009)
10. Lescha, M.F., Hancock, P.A.: Driving performance during concurrent cell-phone use: are drivers aware of their performance decrements? Accident Analysis and Prevention 36, 471–480 (2004)
11. Liu, Y.: Effects of taiwan in-vehicle cellular audio phone system on driving performance. Safety Science 41, 531–542 (2003)
12. Laberge, J., Scialfa, C., White, C., Caird, J.: The Effect of Passenger and Cellular Phone Conversations on Driver Distraction. In: Transportation Research Record, pp. 109–116. Transportation Research Board, Washington (2004)
13. Mccartt, A.T., Hellinga, L.A., Bratiman, K.A.: Cell Phones and Driving: Review of Research. Traffic Injury Prevention 7(2), 89–106 (2006)
14. McKnight, A.J., McKnight, A.S.: The effect of cellular phone use upon driver attention. Accident Analysis and Prevention 25, 259–265 (1993)
15. Miller, G.V.F., Travers, C.J.: Ethnicity and the experience of work: Job stress and satisfaction of minority ethnic teachers in the UK. International Review of Psychiatry 17(5), 317–327 (2005), doi:10.1080/09540260500238470
16. Rakauskas, M.E., Gugerty, L.J., Ward, N.J.: Effects of naturalistic cell phone conversations on driving performance. Journal of Safety Research 35, 453–464 (2004)
17. Shinar, D., Tractinsky, N., Compton, R.: Effects of Practice with Auditory Distraction in Simulated Driving. In: Transportation Research Board 81st Annual Meeting Compendium of Papers (CD-ROM), Transportation Research Board, Washington, DC (2002)
18. Strayer, D.L., Drews, F.A., Crouch, D.J., Johnston, W.A.: Why Do Cell Phone Conversations Interfere with Driving? In: Walker, W.R., Herrmann, D. (eds.) Cognitive Technology: Transforming Thought and Society. McFarland and Company Inc., Jefferson (2005)
19. Strayer, D.L., Drews, F.A.: Profiles in driver distraction: Effects of cell phone conversations on younger and older drivers. Human Factors 46, 640–649 (2004)

20. Strayer, D.L., Drews, F.A., Crouch, D.J.: A Comparison of the Cell Phone Driver and the Drunk Driver. Human Factors 48(2), 381–391 (2006)
21. Strayer, D.L., Johnston, W.A.: Driven to distraction: dual task studies of simulated driving and conversing on a cellular telephone. Psychological Science 12, 462–466 (2001)
22. Wood, C., Torkkola, K., Kundalkar, S.: Using Driver's Speech to Detect Cognitive Workload. In: 9th Conference on Speech and Computer (SPECOM 2004). International Speech Communication Association Press, France (2004)

An Auditory Display to Convey Urgency Information in Industrial Control Rooms

Anna Sirkka , Johan Fagerlönn, Stefan Lindberg, and Ronja Frimalm

Interactive Institute Swedish ICT, Acusticum 4,
94128 Piteå, Sweden
{anna.sirkka,johan.fagerlonn,stefan.lindberg,
ronja.frimalm}@tii.se

Abstract. Auditory warning signals are common features in industrial control rooms. Finding sound signals that convey higher degrees of urgency while keeping the potential for annoyance low is challenging. In the present study, evaluations were performed on four different types of auditory displays. The displays were all designed to convey three levels of urgency. The examination focused on the following questions: (1) "How reliably can the operators identify the three levels of urgency?" and (2) "How annoying do the operators find the sound signals?". Fourteen operators participated in the study. For every signal within each auditory display, the participants were asked to rate the level of urgency and annoyance. The results show that one can design auditory displays that employ appropriate urgency mapping while the perceived annoyance is kept at a low level. The work also suggests that involving the end users in the design process could be advantageous.

Keywords: auditory display, control room, urgency, annoyance, warnings.

1 Introduction

Auditory warning signals are common features in many user environments, including vehicles, clinical facilities and industrial control rooms. Sound has certain advantages over other modes of interaction, especially in critical situations that require immediate attention. Salient auditory cues catch our attention; because hearing is omnidirectional, the sound can be perceived from any direction and wherever the operator has visual focus. Sound can provide information without adding visual load, which can be beneficial in demanding situations that require visual information processing (e.g., monitoring several process parameters on a display).

Nonetheless, the implementation of auditory warning signals is frequently careless. Edworthy [1] reports that the sound signals are too loud, too numerous and too confusing. Other authors have discussed the inappropriate use of auditory signals in a range of user contexts, including airplane cockpits and medical operating rooms [2, 3]. In this study, we focus on the design of auditory warnings for industrial control rooms. The main objective is to develop auditory displays that assist operators effectively, while contributing to a better overall work environment.

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 533–544, 2014.

1.1 Urgency Mapping

Auditory warnings are designed to convey a sense of urgency. The term "urgency mapping" has been defined as matching the perceived urgency of a warning with the urgency of the threatening situation [4]. Appropriate urgency mapping is preferable, as it can help operators prioritize new information and minimize confusion. Inappropriate mapping may, however, have the opposite effect and potentially increase the workload. Therefore, a holistic approach in which all warnings in the operators' environment are considered according to the urgency mapping principle is essential to warning design. This approach is in accordance with the recommendations of the Engineering Equipment and Materials Users' Association (EEMUA), which states that an integrated design should be developed for all auditory warnings in a control room and that the operators' ability to identify the priority of the alarms is desirable [5].

Previous research has established that perceived urgency depends on the fundamental properties of the sound, including several spectral and temporal parameters [6-8]. By manipulating parameters such as speed and frequency content, a designer can systematically change the perceived urgency of the sound. Undoubtedly, learned associations could potentially "override" these mappings [9, 10]. However, considering the gains in learning time and reduced risk of confusion, adapting the physical characteristics from the very start is preferable.

1.2 Annoyance

Annoyance is an important characteristic to consider when implementing auditory warnings in any user context. In accordance with emotion regulation theory [11], operators may try to avoid experiencing the negative emotions associated with the sound simply by avoiding the sound. Considering that sound is omnidirectional and difficult to ignore, the only way to avoid the sound may be to turn the sound level down or to disable the function entirely. For instance, it has been reported that auditory warnings are frequently turned off in anesthetic operating rooms because of their unpleasant properties [3]. Furthermore, Wiese and Lee [12] reported that the annoyance of auditory warnings could be associated with increased workload levels.

There are many reasons why a sound can become an annoyance. Previous research shows that annoyance can be predicted based on physical and psychoacoustic parameters, such as loudness, sharpness duration and tonality [13-15]. For auditory warnings, it has been reported that increasing the urgency of the signal also increases the perceived annoyance [12, 16, 17]. Therefore, finding sounds that convey higher degrees of urgency while keeping the potential for annoyance low is challenging.

1.3 Design of Warnings for a Control Room

Warnings in different user contexts and situations demand different types of responses. For instance, collision warnings presented in a vehicle require an immediate response (e.g., the driver brakes). We argue that when urgent situations

occur in a complex control room environment, it is generally essential that the operator remain calm and focus on solving the problem. The designers should attempt to find solutions that inform and guide the operator effectively and reliably while minimizing annoyance and disturbance. However, as described above, this task is challenging for the designer.

Although auditory warnings could be designed based purely on previous research results, in the present study, a user-centered design process is employed to find solutions that are more appropriate. Prior research provides an understanding of the parameters that influence urgency, annoyance, and distinguishability. However, the operators may also contribute knowledge regarding their work context, the type of urgent situations that can occur, and the task that needs to be performed. This additional insight can assist the designer in adapting the sounds to make them suitable and more tolerable in the work context.

1.4 Aim

In the present study, evaluations were performed on four types of auditory displays designed to assist operators in industrial control rooms. Each display was designed to convey three levels of urgency (low, medium and high). The examination focused on the following questions.

1. How reliably can the operators identify the three levels of urgency?
2. How annoying do the operators find the sounds?

Two of the evaluated concepts are referred to as Design 1 and Design 2. These auditory displays were designed with operators in a user-centered design process and were compared with two baseline displays to gain insight into the appropriateness of the solutions. Baseline 1 is currently in production and is delivered, along with solutions, from a control system manufacturer. Baseline 2 conveys different levels of urgency mainly by manipulating the frequency content of the sound.

2 Method

2.1 Participants

Fourteen control room operators, 1 female and 13 males, participated in the study. All subjects were employees of Smurfit Kappa Kraftliner Piteå. The mean age of the subjects was 46 years (SD 10). The mean experience of the subjects as operators was 22 years (SD 12). All subjects participated voluntarily. None of the subjects reported any hearing disorders relevant to the study. Originally, 15 operators performed the test but due to ambiguous answers, data from one test were excluded from analysis.

2.2 Apparatus

The test took place in two different control rooms at Smurfit Kappa Kraftliner Piteå. Sounds were reproduced for the subjects using headphones (Philips HP890). The sounds were triggered through an application developed in Java and the test interface was presented on a laptop (Apple MacBook).

2.3 Display Concepts

Baseline 1. The Baseline 1 display uses a set of sound signals that is delivered along with solutions from a control system manufacturer. The signals are supposed to represent three levels of urgency. The signals are all abstract tonal sound signals, but they are quite different in character. The signals are used in their original form, i.e., as acquired from the manufacturer. The low-level warning consists of four tonal sounds, each with a length of approximately 580 ms and separated by approximately 580 ms of silence. The total length of the warning is approximately 4000 ms. The medium-level signal consists of an approximately 2600-ms-long tonal sound. The high-level sound consists of two interpolating 250 ms tones with no separation (the first signal is only 75 ms, but whether the signal is like that when implemented in the system is not clear). Figure 1 shows the FFT vs. time analysis for each signal. The sound levels presented are not the absolute values (levels as perceived by the participants). However, the relative differences are correct.

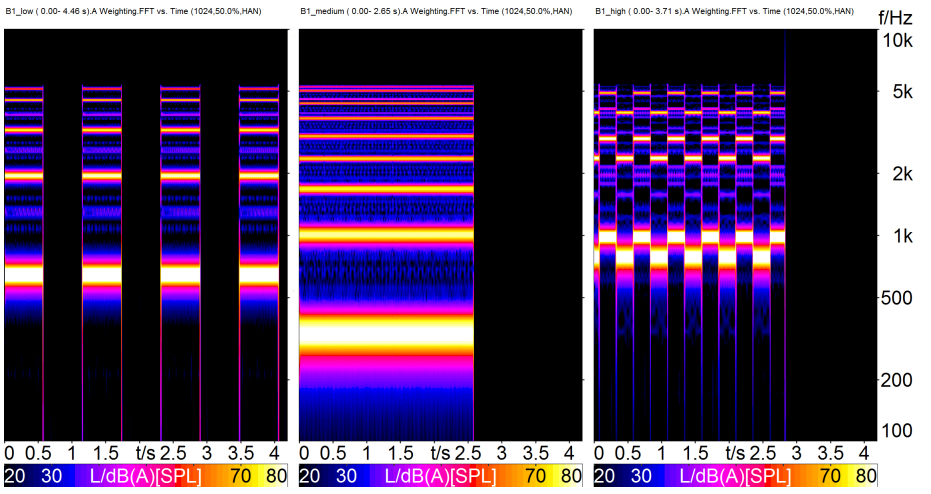


Fig. 1. FFT vs. time for the Baseline 1 sound signals

Baseline 2. Previous research has shown that increasing the fundamental frequency [6] or the amount of inharmonicity or dissonance [7, 18] of a sound increases the perceived urgency. The Baseline 2 display uses a combination of these parameters. The temporal properties were the same for all three signals. The signals consisted of

five tonal sounds, each with a length of approximately 140 ms and separated by approximately 80 ms of silence (the last pause was approximately 50 ms making the signal sound like it was accelerating). The signals were constructed using the following sine tones: low urgency: 300 Hz, medium urgency: 300 Hz + 900 Hz + 3450 Hz, high urgency: 300 Hz + 2450 Hz + 2550 Hz + 3450 Hz + 3513 Hz. There was a slight difference in perceived loudness between the signals, with the low-urgency cue having the lowest level, followed by the medium-urgency and the high-urgency signal. Figure 2 shows the FFT vs. time analysis for each signal within Baseline 2. The sound levels presented are not the absolute values, but the relative differences are correct.

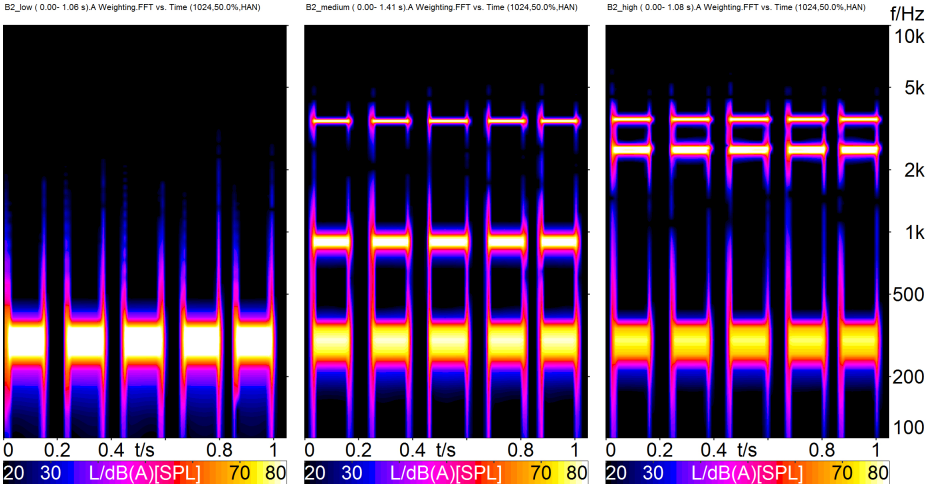


Fig. 2. FFT vs. time for the Baseline 2 sound signals

Design 1 and 2. Previous research has established that the number of repetitions, the speed, and the length of an auditory warning each affect the perceived urgency. Moreover, it has been suggested that the number of repetitions and the speed of a warning are more powerful in inducing changes in perceived urgency than a length change [19].

The Design 1 and 2 displays use a combination of the above-mentioned parameters. The displays have essentially the same temporal structure, where the low-urgency signals consist of one tonal sound with a length of approximately 500 ms. The medium-urgency signals consists of two tonal sounds (total length: 1200 ms) and the high-urgency signals are composed of three tones that are repeated twice (total length: 2300 ms). The signals were subjectively assessed and adjusted in order to have approximately the same sound level.

The following notes comprise the Design 1 signals: low urgency: C4 (261.63 Hz), medium urgency: C4 (261.63 Hz) + D4 (293.66 Hz), high urgency: C4 (261.63 Hz) + D4 (293.66 Hz) + E4 (329.63 Hz). Figure 3 shows the FFT vs. time analysis for each signal. The sound levels presented are not the absolute values.

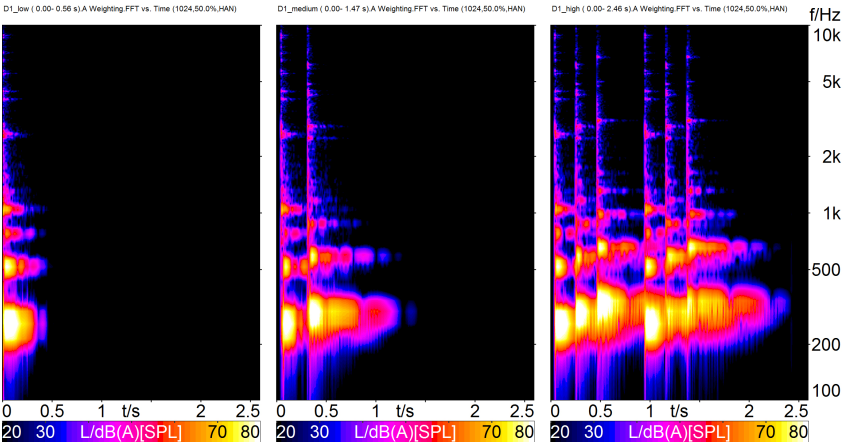


Fig. 3. FFT vs. time for the Design 1 sound signals

Design 2 uses the following notes: low urgency: C4 (261.63 Hz), medium urgency: C4 (261.63 Hz) + F4 (349.23 Hz), high urgency: C4 (261.63 Hz) + F4 (349.23 Hz) + G4 (392.00 Hz). Figure 4 shows the FFT vs. time analysis for each signal. As with the previous analyses, the sound levels presented are not the absolute values.

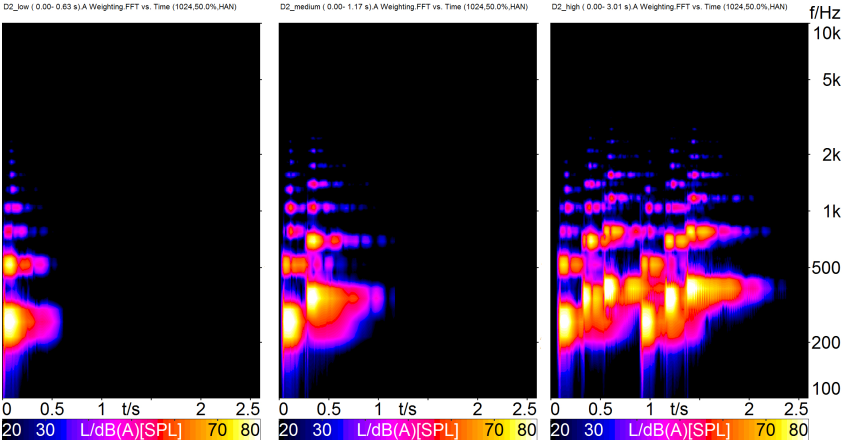


Fig. 4. FFT vs. time for the Design 2 sound signals

The character of the signals in Design 1 and 2 is based on acoustic musical instruments. Soft attacks, long decays, and natural harmonics were used with the aim of creating pleasant sounding signals. The signals of the Design 1 display have a marimba-like sound, and the signals in the Design 2 display are composed of a soft-sounding synth sound.

The Design 1 and 2 displays were developed through a user-driven design process that aimed to develop appropriate alarm sounds for control room environments.

Twenty-four control room operators participated in the process. All subjects worked in the same control room at Smurfit Kappa Kraftliner Piteå. Operators participating in the design process were not used as test subjects during the listening test.

The process comprised an initial workshop followed by approximately ten design iterations. Each iteration consisted of two steps: the development of a design proposal and a user interaction. During a user interaction, a sound design proposal was presented to three to six operators. The design was discussed, and the feedback provided from the operators set the basis for the development of a refined design. Each interaction took place in the operators' own working environment, i.e., in a control room at Smurfit Kappa Kraftliner Piteå, and lasted for approximately half an hour.

2.4 Procedure

The subjects were seated at a table in front of a laptop computer in a remote part of the control room and received written instructions and a questionnaire. Both the instructions and the questionnaire were written in Swedish. The duration of the test was approximately 15 minutes.

The test contained four auditory displays. For each display, three sound signals, named A, B, and C, were judged. The participants listened to the sounds through an interface with three buttons representing signals A, B, and C. The perceived sound levels of the four auditory displays were subjectively adjusted to be approximately the same.

The level of the sound signals was subjectively adjusted to be clearly audible in the present background noise. The subjects could listen to the sound signals of each auditory display as many times as they wished.

To reduce order effects it was desirable that each display was presented first an equal number of times. Therefore the first display to be assessed was specifically chosen prior to testing. The order of the following three displays was randomized, as were the signals within each display. Due to drop-outs the Baseline 2 display initiated a test only twice. The other three displays initiated a test four times each.

For each auditory display, the participants were asked to rate the level of urgency on a stepless scale ranging from "low" to "high". The Swedish word "prioritet" is used by the operators themselves to grade warnings and was used in this study to represent urgency. Similarly, the operators were asked to rate the level of annoyance on a scale ranging from "not at all" to "much". There was also an open comments section for each auditory display.

2.5 Dependent Variables

Ratings of urgency and annoyance are two dependent variables of the evaluation. Additionally, the number of subjects that successfully estimated the correct order of urgency (for all three sounds in the display) is a dependent variable.

3 Results

3.1 Urgency

Figure 5 shows the results of the urgency rating for the sound signals in each display type. All displays, except Baseline 1, resulted in mean ratings indicating an appropriate urgency mapping.

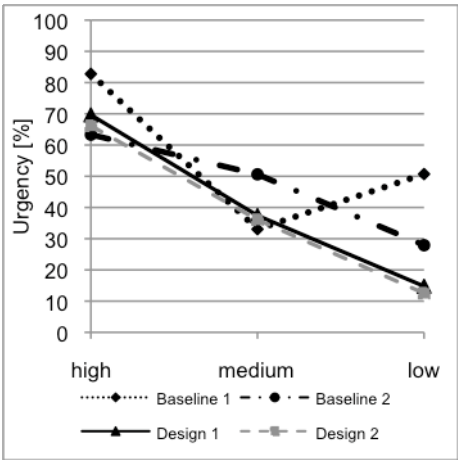


Fig. 5. Mean urgency ratings (100%=high urgency)

A one-way ANOVA, followed by post-hoc analysis (Tukey's HSD test), showed that, for Design 1 and 2, each signal in the display differed significantly from the other two (α=0.05). For Baseline 1, the differences between sound signals were significant, but the low-level signal was rated more urgent than the medium-level signal. For Baseline 2, the difference between the medium- and high-level signals was not statistically significant.

Figure 6 shows how many subjects successfully estimated the (intended) urgency levels for the auditory displays. For Design 1, all participants rated the urgency levels correctly, while for Baseline 1, only two subjects rated the urgency levels as intended.

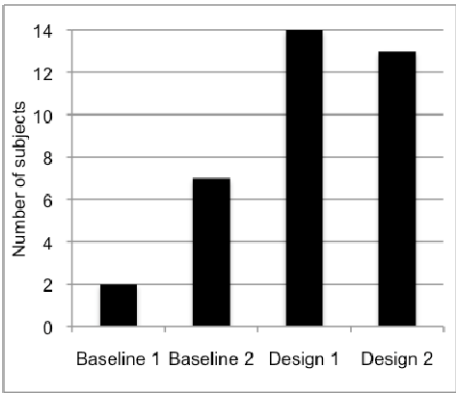


Fig. 6. The number of successful urgency estimations

3.2 Annoyance

Figure 7 shows the mean annoyance ratings for each display type. As expected, sound signals were rated more annoying for the higher urgency levels. However, the signals in Design 1 and 2 displays received low or intermediate mean annoyance scores, while the signals in the baseline displays received intermediate or high mean scores.

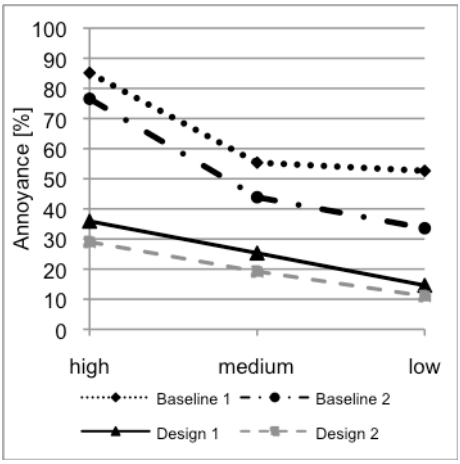


Fig. 7. Mean annoyance ratings (100%=much)

4 Follow-Up Study for Baseline 1

During the evaluation, it was realized that the lengths of the Baseline 1 signals varied. The length of an auditory warning may affect the perceived urgency [19]. For Baseline 1, the low-urgency signal was rated more urgent than the medium-urgency signal. As the low-urgency signal was approximately 1400 ms longer than the

medium-urgency signal (low=4000 ms, medium=2600 ms), a follow-up study was conducted in order to find out whether this difference in length could be the reason for the incorrect urgency mapping.

Fourteen subjects (not control room operators), 6 female and 8 males, participated in the follow-up study. The mean age of the subjects was 35 years (SD 14). All subjects participated voluntarily. None of the subjects reported any hearing disorders that would be of relevance for the study.

The test took place in a desktop environment and the procedure was the same as in the main study. The signals within the display were randomized and participants were asked to rate the level of urgency of the Baseline 1 signals. Consequently, ratings of urgency constituted one dependent variable of the evaluation. Additionally, as in the main study, the number of subjects that successfully estimated the correct order of urgency constituted a dependent variable.

The mean values for the low-urgency and medium-urgency signals were 41.4 % and 43.0 %, respectively (where 100% is representing high urgency). A one-way ANOVA, followed by post-hoc analysis (Tukey's HSD test) showed no significant differences between the two signals. Apart from that, the differences were significant ($p<0.01$). Five subjects rated the urgency levels as intended.

5 Discussion

How reliably can the operators identify the three levels of urgency?

The results support that, for the Design 1 and 2 displays, operators can reliably identify the three levels of urgency. For Design 1, all participants judged the urgency levels as intended. The sounds used a combination of spectral and temporal parameters to express different urgency levels. We cannot draw any conclusions regarding individual parameters and their impact on perception, but taken together, the combination of parameters and their levels resulted in two very promising display designs.

In the Design 1 and 2 displays, the high-urgency sounds received relatively low mean scores. These results are not surprising considering the non-intrusive design of the signals. Auditory warnings can definitely be shaped to sound more urgent by manipulating the sound parameters to more extreme levels. However, we argue that it is essential that the operators can easily and reliably distinguish between urgency levels. However, low levels of perceived urgency may be of importance and cause confusion in user environments where the warnings occur less frequently and where operators do not have the chance to learn the meanings of the sound signals.

The results of the present work show that the signals in Baseline 1 (which is currently delivered by a control system manufacturer) employ an inappropriate urgency mapping. The main study indicates that switching the low- and medium-urgency signals would result in a better mapping. The follow-up study (which tested sounds of equal length) did not support that either of these two signals is perceived as more urgent than the other. For the Baseline 2 display, the differences in scores between signals were not particularly large and the difference between the high- and

medium-urgency signals was not statistically significant. The combination of sound parameters and the selected levels seem to be insufficient in making the sound signals distinguishable in terms of urgency.

How annoying do the operators find the sound signals?

The operators judged none of the signals in the Design 1 and 2 displays to be particularly annoying. Even the high-urgency signals received low to intermediate mean scores as opposed to the high-urgency signals in the Baseline 1 and 2 displays, which received very high annoyance ratings. The rather low annoyance ratings observed for the Design 1 and 2 displays support the appropriateness of these solutions. We cannot draw any conclusions regarding individual sound parameters and their impacts on annoyance. Thus, we cannot make any statements regarding exactly what made the signals in the Baseline 1 and 2 displays more annoying.

A system manufacturer currently delivers the sound signals used in Baseline 1. Conclusions regarding the appropriateness of these signals based on the annoyance ratings observed in the present study should be made with caution. One parameter that might influence annoyance levels is the duration of the signal (a longer duration might be more annoying). The sound signals used in Baseline 1 had longer durations than the signals in the other displays (which may have contributed to higher annoyance levels). However, the extent to which the selected durations for the Baseline 1 signals represent "typical" durations when the signals are implemented in real control room settings was not investigated.

In conclusion, the results support Display 1 and 2 as appropriate auditory displays to convey urgency information to control room operators. The work also support that auditory warning displays can be designed to employ appropriate urgency mapping while keeping the perceived annoyance of the sound signals at a low level. In real implementations, the annoyance of alarms may depend on a range of factors (alarm frequency, false alarm frequencies, etc.) that were not investigated in the present study. Still, the results support that it is worthwhile for system and sound designers to try to lower the perceived annoyance levels of control room alarm sounds. A sound signal that is both effective and has non-annoying characteristics is more likely to become tolerable. Finally, the present work suggests that involving the end users in the design process could be advantageous in reaching successful auditory display solutions.

Acknowledgments. Research presented in this paper was funded by EU:s Structural Funds, the County Administrative Board of Norrbotten, Piteå Municipality, Skellefteå Municipality, and RISE.

References

1. Edworthy, J.: The Design and Implementation of Non-verbal Auditory Warnings. Applied Ergonomics 25(4), 202–210 (1994)
2. Ulfvengren, P.: Design of Natural Warning Sounds. In: Scavone, G.P. (ed.) Proceedings of the 13th International Conference on Auditory Display, Schulich School of Music, McGill University, Montreal, pp. 146–153 (2007)

3. Block, F.E., Nuutinen, L., Ballast, B.: Optimization of Alarms: A Study on Alarm Limits, Alarm Sounds, and False Alarms, Intended to Reduce Annoyance. Journal of Clinical Monitoring and Computing 15, 75–83 (1999)
4. Edworthy, J., Adams, A.: Warning Design: A Research Perspective. Taylor & Francis Ltd., London (1996)
5. Engineering Equipment and Materials Users' Association (EEMUA): Alarm Systems – A Guide to Design, Management and Procurement. EEMUA, London (2007)
6. Edworthy, J., Loxley, S., Dennis, I.: Improving Auditory Warning Design: Relationship between Warning Sound Parameters and Perceived Urgency. Human Factors 33(2), 205–231 (1991)
7. Hellier, E.J., Edworthy, J., Dennis, I.: Improving Auditory Warning Design: Quantifying and Predicting the Effects of Different Warning Parameters on Perceived Urgency. Human Factors 35(4), 693–706 (1993)
8. Haas, E.C., Edworthy, J.: Designing Urgency Into Auditory Warnings Using Pitch, Speed and Loudness. Computing & Control Engineering Journal 7(4), 193–198 (1996)
9. Guillaume, A., Drake, C., Rivenez, M., Pellieux, L., Chastres, V.: Perception of Urgency and Alarm Design. In: Nakatsu, R., Kawahara, H. (eds.) Proceedings of the 8th International Conference on Auditory Display, pp. 357–361. Advanced Telecommunications Research Institute, Kyoto (2002)
10. Burt, J.L., Bartolome, D.S., Burdette, D.W., Comstoc, J.R.: A Psychophysiological Evaluation of the Perceived Urgency of Warning Signals. Ergonomics 38(11), 2327–2340 (1995)
11. Gross, J.J.: Emotion Regulation in Adulthood: Timing is Everything. Current Directions in Psychological Science 10(6), 214–219 (2001)
12. Wiese, E.E., Lee, J.D.: Auditory Alerts for In-vehicle Information Systems: The Effects of Temporal Conflict and Sound Parameters on Driver Attitudes and Performance. Ergonomics 47(9), 965–986 (2004)
13. Hiramatsu, K., Takagi, K., Yamamoto, T., Ikeno, J.: The Effect of Sound Duration on Annoyance. Journal of Sound and Vibration 59(4), 511–520 (1978)
14. Khan, M.S., Johansson, Ö., Sundback, U.: Development of an Annoyance Index for Heavy-duty Diesel Engine Noise Using Multivariate Analysis. Noise Control Engineering Journal 45(4), 157–167 (1997)
15. Landström, U., Åkerlund, E., Kjellberg, A., Tesarz, M.: Exposure Levels, Tonal Components and Noise Annoyance in Working Environments. Environment International 21(3), 265–275 (1995)
16. Marshall, D.C., Lee, J.D., Austria, P.A.: Alerts for In-vehicle Information Systems: Annoyance, Urgency and Appropriateness. Human Factors 49(1), 145–157 (2007)
17. Tan, A.K., Lerner, N.D.: Multiple Attribute Evaluation of Auditory Warning Signals for In-vehicle Crash Avoidance Systems. Technical Report, National Highway Traffic Safety Administration (1995)
18. Russo, F.A., Lantz, M.E., English, G.W., Cuddy, L.L.: Increasing Effectiveness of Train Horns Without Increasing Intensity. In: Brazil, E., Shinn-Cunningham, B. (eds.) Proceedings of the 9th International Conference on Auditory Display, pp. 51–54. Boston University Publications Production Department, Boston (2003)
19. Hellier, E., Edworthy, J.: Quantifying the Perceived Urgency of Auditory Warnings. Canadian Acoustics 14(4), 3–11 (1989)

Hierarchical Task Analysis of a Synthetic Aperture Radar Analysis Process

Susan Stevens-Adams, Kerstan Cole, and Laura McNamara

Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185, USA
{smsteve,kscole,lamcnam}@sandia.gov

Abstract. Imagery analysts are given the difficult task of determining, post-hoc, if particular events of importance had occurred, employing Synthetic Aperture Radar (SAR) images, written reports and PowerPoint presentations to make their decision. We were asked to evaluate the current system analysis process and make recommendations for a future temporal geospatial analysis prototype that is envisioned to allow analysts to quickly search for temporal and spatial relationships between image-derived features. As such, we conducted a Hierarchical task analysis (HTA; [3], [6]) to understand the analysts' tasks and subtasks. We also implemented a timeline analysis and workload assessment [4] to better understand which tasks were the most time-consuming and perceived as the most effortful. Our results gave the team clear recommendations and requirements for a prototype.

Keywords: Hierarchical Task Analysis, Synthetic Aperture Radar, timeline analysis, workload assessment.

1 Introduction

Imagery Analysis refers to the perceptual and cognitive work of detecting and identifying features of interest in two-dimensional renderings of remotely sensed data. Analysis of imagery is a core activity in many fields, from radiology to military operations planning to civilian intelligence production. Over the past two decades, computers have emerged as the primary vehicle for rendering and displaying imagery, but its review and interpretation remains primarily a human activity. Indeed, in government military and intelligence activities, imagery analysts typically undergo months of rigorous training to acquire the skills that enable reliable, accurate identification of features in products derived from a broad spectrum of sensing systems [5].

Yet the human 'eyes-on-imagery' analysis paradigm is increasingly strained by the sheer scale of digital image datasets in commercial, government and academic domains. For example, over the past decade, the United States' military and intelligence communities have invested significantly in remote sensing technologies with the goal of increasing the quality and quantity of information to support both tactical and strategic decision-making. The investments have been wildly successful, providing military and intelligence functions with impressive remote sensing capabilities that have, in turn, swamped those same functions with unprecedented

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 545–554, 2014.

amounts of sensor data. In response, government agencies are seeking capabilities that will enable management of the so-called 'data deluge'; i.e., technologies to automate the processing and analysis of sensor datasets, so that military and intelligence personnel are able to realize the expected information value of the terabytes of sensor data that systems are generating.

Importantly, government agencies are not the only entities dealing with a data deluge. Commercial and academic organizations are also seeking technology that will facilitate human processing and evaluation of imagery; e.g., search systems that can retrieve image content without reliance on text tags. Such systems present considerable design challenges: the human visual system is exquisitely capable of recognizing and classifying patterns into contextually-relevant information, an interpretive task that is quite challenging for even the most sophisticated algorithms running on the fastest processors.

Rather than automating the interpretive work of imagery analysis, a better design goal is to identify and address elements of image-related work that are laborious to humans but well-suited for automation. For example, most digital cameras include image processing software that reduces visual noise by minimizing glare or enhancing contrast in an image, making key features more detectable and recognizable to human viewers. In the context of very large image sets – for example, giga- or tera-sized image datasets – one might imagine algorithms that enhance pixel-based temporal or geospatial patterns over multiple sensor datasets, in ways that enable people to engage larger, more detailed, even heterogeneous data resources, without introducing extraneous cognitive, motor, and/or perceptual load.

In this paper, we discuss the use of Hierarchical Task Analysis (HTA) to support just such a set of algorithmic design goals [4], [6]. We used HTA and associated methods, described below, to understand how offline analysts work with a particular type of imagery, namely the pixel-based renderings generated by of Synthetic Aperture Radar (SAR) systems. This work is part of a larger informatics effort entitled PANTHER - **P**attern **AN**alytics **T**o support **H**igh-performance **E**xploitation and **R**easoning. PANTHER is a three-year research and development project that that brings together statistics, graph algorithms, software engineering, visualization and human factors to automate the extraction and aggregation of key features captured in remotely sensed datasets, with the goal of enabling humans to perform the interpretive work of pattern recognition, classification and contextualization over temporally and geospatially distributed image sets.

First developed over fifty years ago, SAR systems typically comprise a radar and antenna side-mounted on a small aircraft. As the aircraft flies, the radar pulses the ground with millimeter-scale radio waves. Echoes or 'returns' are captured by the antenna, then stored, processed, and rendered for visual inspection as pixelated imagery (see Figure 1 for an example). Because SAR systems can acquire high-resolution imagery in day or night, even in inclement weather, these systems have found use in a wide range of applications including reconnaissance, environmental monitoring and activity detection. Importantly, SAR systems can be used to repeatedly image the same scene over extended time periods. By analyzing differences in the coherence and magnitude of returned signals, it is possible to create imagery that indicates scene changes occurring between radar passes.

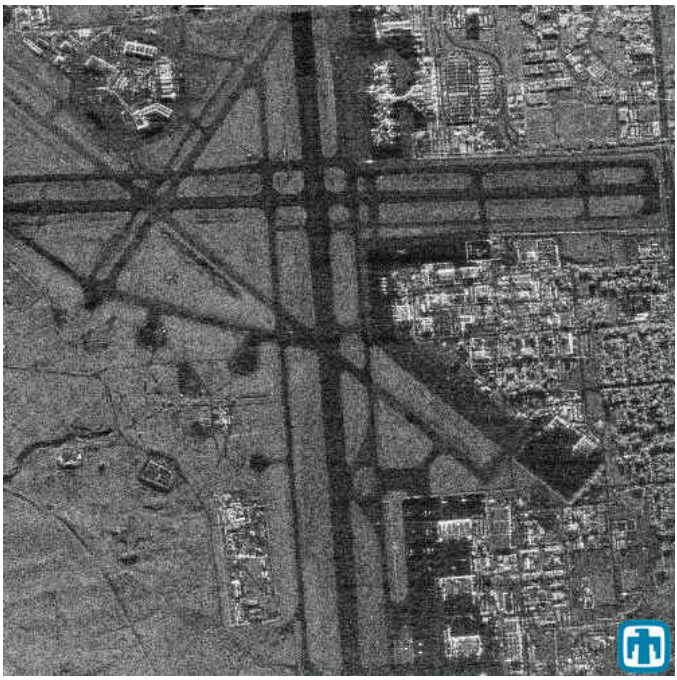


Fig. 1. SAR imagery of the Albuquerque Airport

2 Imagery Analysis in Mongoose

Our design research has focused on SAR imagery analysts associated with "Mongoose," an imaging system used to support both military and civilian field operations. PANTHER computer scientists are partnering with Mongoose stakeholders to develop algorithms that take advantage of the rich coherence and magnitude change information in SAR datasets, to reveal temporal and spatial patterns captured in Mongoose imagery. Studies of Mongoose imagery analysts are providing key information not just about the patterns that imagery analysts are seeking to characterize, but elements of the workflow that offer opportunities to reduce extraneous workload: e.g., manual cut-and-paste actions to associate reporting content from one database with images in another database.

Over the past year, we have studied how imagery analysts at different points of the Mongoose workflow interact with SAR image products to identify and characterize changes in scene content over periods ranging from several hours to multiple days. Mongoose employs imagery analysts at two points in the system workflow: "near-real time" imagery analysts work with the fielded sensor, reviewing and evaluating imagery as it comes from the aircraft. A complementary "offline" process involves reviewing imagery and reports to determine, post-hoc, whether Mongoose teams in the field have accurately characterized trends and events of interest. Despite the fact that this analysis takes place "offline," it is still done under time pressure and with the

knowledge that the feedback may influence the judgments of the field imagery analysts who work in "near-real time"; that is, making rapid trend and event assessments using Mongoose imagery as it comes off the radar.

The HTA described in this paper focuses the work of "offline" imagery analysts. Not only do they review reported events to evaluate the correctness and completeness of field reports, but they frequently augment field reports with additional information that may not be available to the fielded teams. Offline analysts use a variety of forms and presentation templates to capture their analysis, with products populating a Mongoose database for all field events reported during the Mongoose system's lifetime. As we discuss below, the current workflow involves a number of onerous, time-consuming tasks that do not contribute significant content to the analysis products, but which do consume significant attention and energy. By studying the current workflow, our team has made recommendations for a future temporal geospatial analysis prototype that is envisioned to allow analysts to quickly search for temporal and spatial relationships between image-derived features.

2.1 Methodology

Our approach to HTA begins with analysis of the broader work domain methods from Cognitive Work Analysis (CWA). Previous research has suggested that the CWA and HTA methods are complimentary, with CWA being useful to inform the design and implementation of HTA focused on specific tasks in the workspace [3]. Of particular importance was the development of a CWA Abstraction Hierarchy and Decision Ladder, both of which we found very useful in identifying key activities suitable for focused inquiry of HTA. By developing these representations of the work domain and key decision processes, we were able to put the offline analysis process into the context of the broader Mongoose workflow. This workflow begins with imagery analysis tasks at the fielded sensor and ends with population of the offline Mongoose event database mentioned above (see [1] for a description of the work and findings).

In addition to the HTA, we also conducted a Timeline Analysis and Workload Assessment on the tasks derived from the HTA to shed light on which tasks were taking the most amount of time and were most effortful. PANTHER algorithm and software developers have been using our analysis products to identify areas where automation of tasks is most likely to have significant benefits for the analytic community, in terms of reducing time and effort spent on processes that do not contribute to event and trend analysis.

3 Hierarchical Task Analysis (HTA)

Hierarchical Task Analysis involves the study of what an operator is required to do, in terms of actions and/or cognitive processes to achieve a system goal [4]. Three principles govern HTA approaches [6]: a system is described first in terms of its goals; then, the operation can be broken down into sub-operations, each of which is defined by a measured sub-goal. The analysis posits a hierarchical relationship between operations, sub-operations and, by extension, between goals and sub-goals.

We took the framework from Stanton [6] for conducting a HTA and summarize the authors' recommendations for conducting an effective HTA:

1. **Define the purpose of the analysis**. Stanton [6] (hereafter referred to as 'the author') emphasizes the importance of identifying the expected outcomes of an HTA analysis prior to starting data collection. In our case, the purpose of the analysis was to obtain an understanding of the analysts' workflow, the tasks performed and the task resources, with the goal of identifying high-effort, low-learning-payoff tasks and provide recommendations and requirements for analytic software that automated such tasks.
2. **Define the boundaries of the system description.** The author emphasizes that the boundaries of the system will depend upon the purpose. We were interested in the offline analysts' tasks and workflow as a subset of the larger Mongoose workflow. The analysts were required to access multiple databases (some of which were not in-house) and use several in-house computers in order to perform and complete their analysis. We did not address the SAR sensor system nor examined other areas of the larger Mongoose workflow, which is quite extensive.
3. **Try to access a variety of sources of information about the system to be analyzed**. We used multiple sources to gain an understanding of the system. We observed and interviewed two offline imagery analysts (subject matter experts) for 50+ hours. The analysts performed and talked through their daily tasks, which included transferring data from a server, consolidating and reading relevant reports and PowerPoint presentations, looking through SAR imagery and determining what transpired for all events. We also observed, documented and participated in weekly team meetings at which the offline analysts discussed updates and current issues.
4. **Describe the system goals and sub-goals.** The overall aim of the analysis was to derive a sub-goal hierarchy for the tasks. Based on our observations and interviews, we were able to describe the goals and sub-goals and relate these to specific operations and sub-operations.
5. **Try to keep the number of immediate sub-goals under any super-ordinate goal to a small number.** The author suggests keeping the number of immediate sub-goals to between 3 and 10; we kept to this recommendation.
6. **Stop re-describing the sub-goals when you judge the analysis is fit-for-purpose.** Once we obtained an understanding of the offline analysts' tasks and resources, we judged the analysis fit-for-purpose. In this case, our purpose was to identify onerous tasks that could be automated to reduce extraneous workload for Mongoose's offline imagery analysts.
7. **Try to verify the analysis with the subject-matter experts**. The author states that it is important to check the HTA with subject matter experts to verify the completeness of the analysis and help the experts develop a sense of ownership of the analysis. We met with our offline imagery analyst participants throughout our analysis. We engaged them in discussions about the completeness and correctness of our research and representations as iterated to our fit-for-purpose. We used the offline analysts' feedback to refine our representations and recommendations for the PANTHER development team.

8. **Be prepared to revise the analysis**. Based on our evaluation discussions with the offline analysts we revised our analysis multiple times. Only after the experts were in agreement that our representations were complete and correct did we stop iterating and provide our analysis to the PANTHER development team.

4 HTA Analysis

Our HTA revealed that Mongoose's offline analysts performed six distinct tasks under the larger goal of "Analyze and Evaluate Reported Trends and Events" (abbreviated as "Analyze Event" in the hierarchy example in Figure 2). All tasks consisted of multiple subtasks. We also identified offline analysts' decision points, as well as potential errors and sources of bias that we believed to influence their evaluation of field team performance. The six tasks and associated sub-tasks are represented in Figure 2.

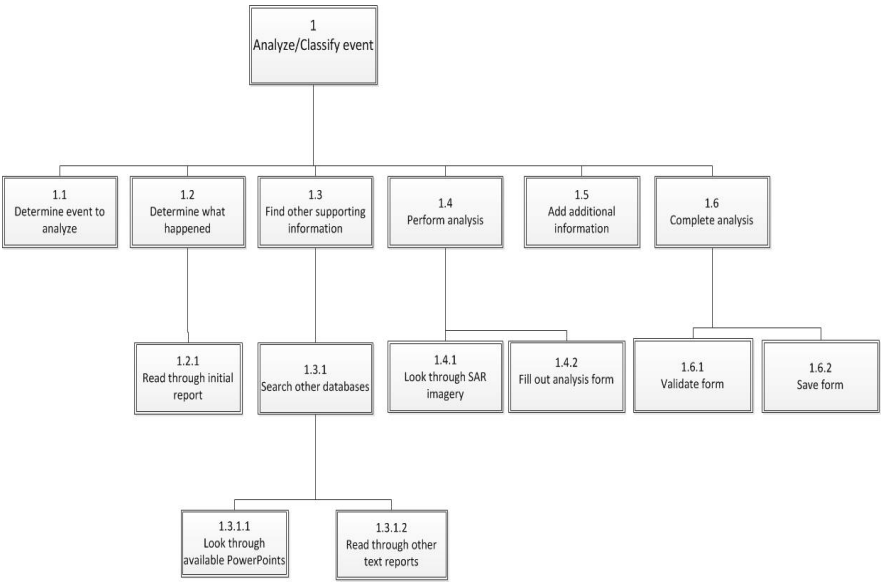


Fig. 2. Hierarchical diagram of the goal of "Analyze Event"

In breaking down the offline analysis process hierarchically, we learned that the offline analysis process draws heavily on the written (text) reports. To our surprise, our study participants spent minimal time visually inspecting the SAR imagery. In fact, analysts tended to make a decision about the correctness and completeness of possible events and trends of interest *prior* to any visual inspection of the imagery data.

Given this task sequence, we wondered if the offline analysts' evaluation might be prone to confirmation bias. Despite the fact that the SAR image data contains significant information about events and trends on the ground, our participants spent minimal time with the imagery, and a significant amount of time reviewing text datasets, such as event reports. This was a surprise to the algorithm developers on the PANTHER team, whose members had assumed that all SAR analysts would rely first on image products, using auxiliary non-image data as a secondary information source. However, the offline analysts pointed out that text datasets contained important contextual metadata that was not easy to extract from the SAR image sets, and which were critical for their analytic work. As a result, the offline analysts had learned to read and select information items from non-imagery sources that contained meta-information about the trends and events under study; the imagery associated with these events and trends was primarily useful in understanding how the field analysts had assembled their reporting, and to illustrate details of the region in which such events and trends had taken place.

5 Additional Analysis

5.1 Timeline Analysis

In order to more quantitatively assess which of the tasks was most time consuming, we implemented a timeline analysis. A timeline analysis is a method of identifying the density of tasks to be performed [4]. We were interested in how long it took the analysts to perform each of the tasks identified in the HTA. As such, we asked three analysts to record how long it took them to perform each of the HTA tasks. We compiled the results and found that, on average, the analysts spent the most amount of time searching other databases for information, reading relevant reports and completing their analysis (see Figure 3). The analysts informed us that, in terms of finding other supporting information (Task 3), they spent roughly the same amount of time on the sub-tasks of reading through other text reports and looking through available PowerPoint slide decks. In terms of performing their analysis (Task 4), the analysts noted that the sub-task of filling out the analyst form was much more time consuming than the sub-task of looking through the SAR imagery.

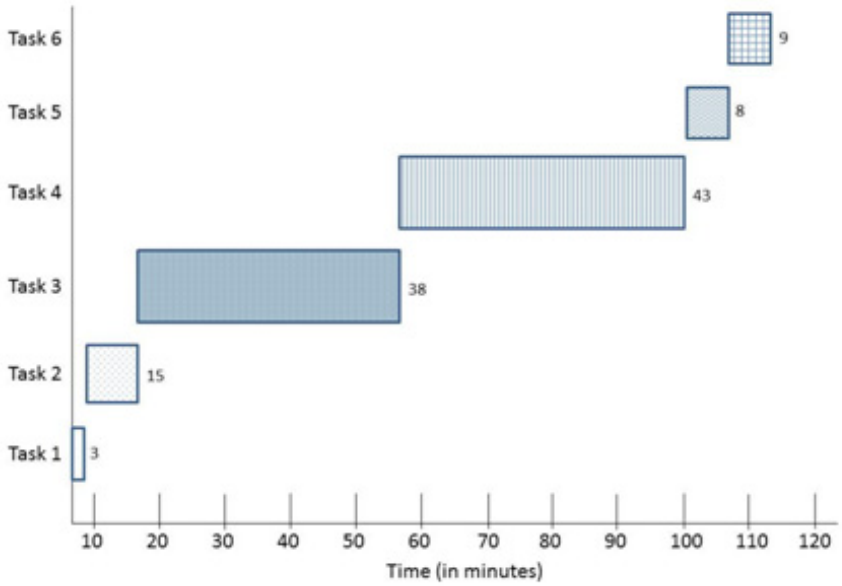


Fig. 3. Timeline assessment of tasks

5.2 Workload Assessment

In order to assess which of the tasks were perceived as the most effortful, we performed a workload assessment via the NASA Task Load Questionnaire (NASA TLX; [2]). After completing each task, the analysts were asked to rate their perceived workload (or effort) on a scale of 1 (very low) to 7 (very high) for five different areas; mental demand, physical demand, temporal demand, effort and frustration and on a scale of 1 (perfect) to 7 (failure) for the area of performance. Three analysts completed the workload assessment and their scores were averaged (see Table 1). None of the tasks were rated as particularly physical or temporal demanding, which is aligned with the fact that the analysts use a simple computer and keyboard to perform their work and are generally given sufficient time to complete their analysis. We found that those tasks that were more rote (Tasks 1 and 6) were rated as the least mentally demanding, the least effortful and the least frustrating. However, those tasks that required more cognitive resources, reasoning and decision making (Tasks 2-5) were considered more mentally demanding, more effortful and the most frustrating.

Table 1. Average NASA TLX ratings for each analysis task

	Mental demand	Physical demand	Temporal demand	Performance	Effort	Frustration
Task 1	1.0	1.0	1.0	1.0	1.0	2.0
Task 2	3.3	1.3	1.3	1.7	2.8	1.7
Task 3	3.3	1.7	1.3	2.0	2.7	2.5
Task 4	3.7	1.7	1.3	1.7	3.0	2.3
Task 5	2.7	1.7	1.3	2.0	2.2	1.7
Task 6	1.3	1.0	1.0	2.0	1.3	1.3

6 Recommendations for Prototype

Based on the results from our analysis, we were able to propose requirements and recommendations for a future system in which the analysts could better utilize SAR imagery. We proposed that a future system be designed intentionally to support a workflow that relies primarily on imagery analysis with auxiliary text descriptions as a secondary or supporting contextual element. The imagery needs to include metadata, as this was very important to the analysts and their analysis. In addition, the system should allow the imagery and metadata to be searchable, through querying. The analysts could benefit from an interactive search capability that would allow them to determine any sort of trend behavior or to pull up any similar events from the past. Finally, it was recommended that the future system auto-populate some of the routine information currently captured in spreadsheet-based forms, since the analysts found the task of filling out the form particularly mundane and tedious. We emphasize that selection of auto-populated information would have to be carefully determined with the analysts' input.

7 Discussion

The imagery analysts in our project were tasked with evaluating the completeness and correctness of reported events and trends of interest, using radar imagery, written reports and PowerPoint presentations. In support of the PANTHER informatics project, our team was asked to evaluate the current system analysis process and make recommendations for a future temporal geospatial analysis prototype that is envisioned to allow analysts to quickly search for temporal and spatial relationships between image-derived features. We used HTA, Timeline Analysis and Workload Assessment to better understand offline analysts' workflow. In regards to the current workflow, the offline analysts focused most of their attention on the written reports and PowerPoint presentations, and spent little time looking at the SAR imagery. This was news to the team leads, as they assumed that the analysts were primarily using the SAR imagery to make their decisions. The analysts informed us, however, that they were not able to easily access the information they needed from the imagery. Thus, they relied on the written reports and PowerPoint presentations for the pertinent

information and only used the imagery to confirm their decision, which could lead to confirmation bias.

Based on our analysis, we were able to make recommendations and requirements for the design of a future system aimed at minimizing the amount of effort required to complete low-value routine operations, such as cutting and pasting text information into spreadsheet forms. In addition, our recommendations identified barriers to the effective offline exploitation of SAR imagery in this particular analytic workflow. As we go forward, we expect to continue interacting with the analytic teams, the PANTHER developers and other stakeholders in an iterative prototyping-and-evaluation process.

Acknowledgments. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. (SAND2014-1407C).

References

1. Cole, K., Stevens-Adams, S., McNamara, L., Ganter, J.: Applying cognitive work analysis to synthetic aperture radar system. In: The Proceedings of the Human Computer Interaction International Conference (2014)
2. Hart, S.G., Straveland, L.E.: Development of the 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)
3. Jenkins, D., Stanton, N., Walker, G., Salmon, P., Young, M.: Creating interoperability between the Hierarchical Task Analysis and the Cognitive Work Analysis Tool. Report from the Human Factors Integration Defence Technology Centre, U.K. (2006)
4. Kirwan, B., Ainsworth, L.K. (eds.): A Guide to Task Analysis. Taylor & Francis, Florida (1992)
5. Richelson, J.T.: The US Intelligence Community. Westview Press, Colorado (2011)
6. Stanton, N.A.: Hierarchical task analysis: Developments, applications and extensions. Applied Ergonomics 37, 55–79 (2006)

Author Index