Improving trainees' performances while under stress using real-time feedback

Iris Cohen

Improving trainees' performances while under stress using real-time feedback

Proefschrift
ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op woensdag 28 oktober 2015 om 15:00 uur

door
Iris COHEN
MSc Toegepast cognitief psycholoog
Universiteit Utrecht, Nederland
geboren te Leidschendam, Nederland

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. M. A. Neerincx
Copromotor: Dr. ir. W. P. Brinkman

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. M. A. Neerincx	Technische Universiteit Delft, promotor
Dr. ir. W. P. Brinkman	Technische Universiteit Delft, copromotor

Onafhankelijke leden:

Prof. dr. C. M. Jonker	Technische Universiteit Delft
Prof. Dr. Ir. P. A. Wieringa	Technische Universiteit Delft
Prof. dr. J.M.C. Schraagen	University Twente
Prof. dr. ir. S.A. Meijer	Kungliga Tekniska Hogskolan
Prof. Dr.-Ing. F. Flemisch	Fraunhofer FKIE

Printed by: Gildeprint
Cover design by: Glenn Windhouwer
Cover photo by: I. Cohen

ISBN 978-94-6186-525-0
An electronic version of this dissertation is available at http://repository.tudelft.nl/

TABLE OF CONTENTS

1. INTRODUCTION

On July 3rd, 1988, the USS Vincennes, a United States warship, intervened when armed Iranian speedboats showed hostility towards European cargo ships. The Vincennes sent a helicopter to investigate the situation. Shots were fired at the helicopter by the speedboats which started a firefight between the armed Iranian boats and the Vincennes (Zatarain, 2008).

During this firefight, the crew on the Vincennes detected an aircraft on their radar and identified it as an attacking F-14 Tomcat fighter jet. In an act of self-defence, they shot down the aircraft with two radar-guided missiles. Regrettably, it turned out to be an Iranian civilian Airbus carrying 290 civilian passengers and crewmembers (Zatarain, 2008).

Later investigations suggest the stress from combat probably caused task fixation and a distorted perception of the available information. This likely led the Vincennes to mistake the Iranian airbus for an attacking fighter jet. They reported that the aircraft was descending, similarly to attacking aircraft. In reality, however, the airplane was climbing up.

Figure 1.1 On July 3rd, 1988, an American naval warship, the USS Vincennes, accidentally shot down a commercial airliner while in combat with armed speedboats on July 3rd 1988 (Zatarain, 2008).

1.1 BACKGROUND

Professionals working as emergency responders or surveillance officers (e.g., police officers, firefighters, paramedics, and military personnel/servicemen and women), and professionals in other high risk and safety sensitive domains, such as astronautics or aviation, frequently encounter uncertain, complex and risky situations (Driskell & Johnston, 1998). Regardless of the cause of the danger (a natural cause such as hurricanes, floods, or earthquakes, or a human cause such as riots, traffic accidents, or hostile attacks as the one described in figure 1.1), the professionals concerned are required to assess the situation and the needs of potential casualties quickly and accurately and respond accordingly.

These hazardous situations, however, cause or inflate stress inducing factors such as time pressure and personal risk to the professionals, which can result in the perception of stress. This happens when individuals believe the demands of the situation exceed their skills and resources, negatively impacting the physiological, psychological, social and behavioural domains (Salas, Driskell, & Hughes, 1996). The way individuals work under stress depends on individual people and their perception of the demands (Kowalski-Trakofler, Vaught, & Scharf, 2003).

The *USS Vincennes* (figure 1.1) experienced a situation that caused high or even extreme levels of stress. The analyses of this incident show there were many and diverse sets of factors affecting the chain of information processes and decision-making processes, at both the team and individual levels. This thesis investigates the possibilities to improve the training of individuals that perform under these stressful conditions. High levels of stress have been shown to negatively affect several cognitive processes and consequently deteriorate the professional's performance (Keinan, Friedland, & Ben-Porath, 1987; Ozel, 2001; Starcke & Brand, 2012). One of the foremost cognitive processes negatively impacted by stress is the process of decision-making (Kerstholt, 1994; Starcke & Brand, 2012). Unfortunately, during crises, flawless decisions need to be made, since the consequences of an erroneous decision can be disastrous, as illustrated in figure 1.1.

The field of research that focusses on decision making in real-life settings is called naturalistic decision making (NDM), or macrocognition (Schraagen, Ormerod, Militello, & Lipshitz, 2012). An important premise in NDM, as first demonstrated by Kahnemen, Slovic, and Tversky (1982), states that decisions are made based on heuristics. Klein, Calderwood, and Clinton-Cirocco (1986) expanded this idea by showing that firefighters do not make decisions by considering different decision possibilities, but by assessing the situation and comparing it to previous experiences. It is therefore not fruitful to provide insight into different decision options to professionals working in stressful, naturalistic circumstances. This is not the natural way of deciding for them.

The aim of this doctoral dissertation is to improve professionals' decisions and performances when they work in risky, stressful situations. To this end, a digital

support tool that focusses on individuals' reactions to stress and provides support during training was created. As described in the previous paragraph, this support was not based on different decision options but on physiological and cognitive effects of stress. The following chapters describe the creation of this tool and investigations into its effectiveness.

1.2. CURRENT TRAINING AND SUPPORT TOOLS

Preparing professionals for crises or disaster scenarios can be accomplished by letting them experience similar situations. This is called learning by experience or stress exposure training (SET) (Cesta, Cortellessa, & Benedictis, 2014; Driskell & Johnston, 1998). This should be done in a safe training environment, for example during scenario-based training using virtual reality (VR) (Driskell & Johnston, 1998; Peeters, Van Den Bosch, Meyer, & Neerincx, 2014; Salas & Cannon-Browers, 2001). VR makes it possible for professionals to experience stressful situations similar to those they might experience in real life, but without any actual threat (Busscher, Vlieger, Ling, & Brinkman, 2011).

On top of a realistic training environment, appropriate feedback or instructions are necessary during training (Mayer, 2004). There are technical devices and systems that help experts make decisions, known as decision support systems or intelligent decision aids (IDA). Early versions of these systems helped the decision-maker pick the right decision based on gathered information (Kontogiannis & Kossiavelou, 1999). As described in Section 1.1, this is not how professionals make decisions in naturalistic environments (Klein et al., 1986). Reason (1987) also argues that the human operator is better at decision-making, especially in novel situations.

Another form of technical systems that help trainees perform better under stress is biofeedback systems. These systems do not focus on formulating a decision for their users. Instead they increase professionals' awareness of their physiological stress reactions, such as an increased heart rate or respiration rhythm. Users that are aware of their physiological reactions to stress are thought to e.g. regain control of their heart rate which leads to a reduction of overall feelings of stress. Although biofeedback has been found to increase performances (Bouchard, Bernier, Boivin, Morin, & Robillard, 2012), it is still unclear if these effects are long-lasting (Raaijmakers et al., 2013).

The current method of stress training could use a more interactive, specific, and personalized approach (Cohn, Weltman, Ratwani, Chartrand, & McCraty, 2010). Several improvements that could create more interactive, specific, and personalized stress training have been proposed. Kontogiannis and Kossiavelou (1999), for instance, believe that support systems should provide insight into event escalation, rather than merely produce a decision. Support systems should indicate when communication strategies should change, or when task allocation of team members need to be adjusted in order to work optimally in the situation (Kontogiannis & Kossiavelou, 1999). Support

systems can also focus on indicating errors. Dörner and Schaub (1994) state that confronting people with their tendencies to err can decrease the number of errors made.

Although support systems or feedback systems themselves are not new, combining biofeedback and more in-depth and real-time feedback, as suggested by Kontogiannis and Kossiavelou (1999) and Dörner and Schaub (1994), is. Such a support tool is created and tested with experiments that are described in this dissertation.

1.3. PROBLEM STATEMENT, HYPOTHESES AND RESEARCH METHODS

1.3.1. PROBLEM STATEMENT

There is a need for practical support for professionals that work under stressful conditions (see Section 1.1). Digital support tools should focus on individuals' reactions to stress. Technology that helps individuals control their physiological reactions to stress is available. This can be stress caused by events in real-life, but also by virtual environments, as they can also evoke stress in the individuals (see Section 1.2). Less attention has been given to the cognitive and affective reactions stress causes. Suggestions were made (Dörner & Schaub, 1994; Kontogiannis & Kossiavelou, 1999) about focusing support tools on error tendencies or opportunities to switch strategies. However, there is a lack of empirical evidence of the effectiveness of such support tools. This dissertation aims to change that.

1.3.2. HYPOTHESES

The overall aim of this doctoral dissertation is to improve professionals' decision-making processes and performances when they work in hazardous and stressful situations. It is necessary to decrease the negative effects of stress on professionals' performances, for instance with the help of a (technical) support tool that can be used in training settings. The aim can be translated into the following main hypothesis:

"A real-time feedback system improves the performances of trainees' working in stressful environments"

Four research questions were established to increase insight into the main hypothesis. The studies described in this dissertation focus on these research questions:

- Which aspects of the work content influence the cognitive and affective factors of cognitive task performance?
- Can work content and cognitive and affective factors measured in real-time predict trainees' performance in real-time?
- Do trainees' task performances improve by providing real-time predictive feedback during stressful events?
- What type of real-time feedback improves task performances in stressful

scenarios?

Feedback is defined in this dissertation as returning (parts of) the output of a certain process, in this case the process of decision-making while working under stress. Following this definition, a "*feedback system*" is a system that collects or calculates (parts of) the output of a certain process, and returns it to its user.

1.3.3. RESEARCH APPROACH

Question 1 and 2 focus on modelling the process of decision-making under stress. To accomplish this, literature about the process of working and performing under stress was consulted, and an initial model was established (Chapter 2). This model was tested by fitting it to data from a simulated Mars mission (Chapter 3). Results showed that work content factors (e.g., the different task goals and demands of the different computer applications and mission phases) influence astronauts' cognitive and affective processes. Next, the model was fitted to data from a simulated Navy mission (Chapter 4). Models were created to predict two types of performances based on the variables from the model. The performances that could now be predicted were: performance scores rated by experts, and chances of making specific errors during a task. The predictive models from Chapter 4 were implemented into a feedback system that calculated predictive performance values in real time (Chapter 5). The third and fourth research questions focus on the effectiveness of this feedback system. Chapter 6, in which Navy students received feedback in a scenario-based virtual environment, shows an improvement in performances. These findings combined led to a more detailed examination of the feedback, and all the possible combinations of different feedback types provided by the system (Chapter 7).

1.4. DISSERTATION OUTLINE

This dissertation is divided in two parts, as can be seen in figure 1.2. The first part contains three chapters that focus on the establishment and refinement of a model that describes the process of working and performing under stress. The second part, also consisting of three chapters, focuses on evaluating a feedback system based on the variables from the model.

As mentioned earlier in this chapter, the first step in this thesis was to model the process of working and deciding under stress. While the literature provides several models that explain the process of decision-making, Chapter 2 proposes the COgnitive Performance and Error (COPE) model. The COPE model focusses on the effects of acute stress on performances, based on the influence of stress on cognitive and affective factors. Stress is considered acute when it has a novel, sudden, and intense onset (Salas et al., 1996), while prolonged stress is the opposite. The *USS Vincennes*' situation is considered to have evoked acute stress. In Chapter 2, a graphical presentation of this

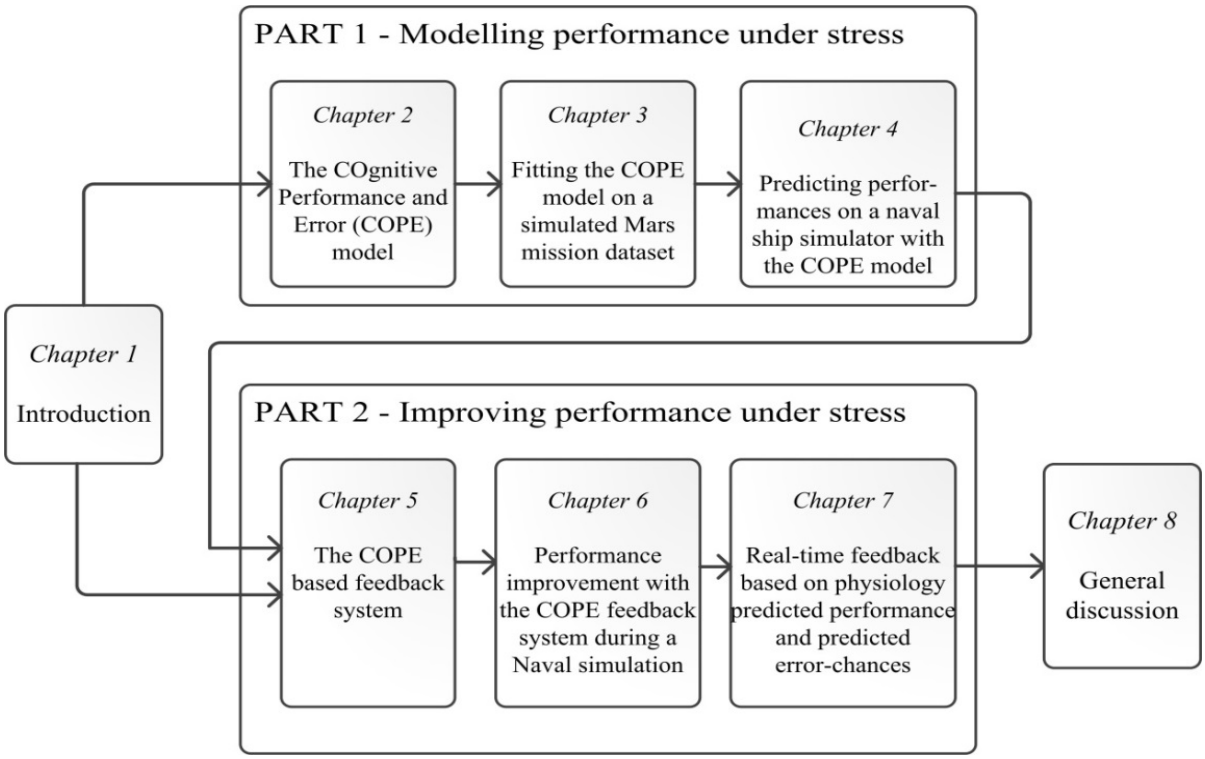


Figure 1.2. Graphical presentation of the thesis outline.

model is provided, and the variables and their relationships are underpinned with literature.

In Chapter 3, a simulated 520-days Mars mission is described. During this long-term isolation mission, participants performed different tasks while rating their cognitive and emotional state. Data from this simulated mission was fitted to the COPE model. By doing so, influences of the work content on cognitive and affective processes and on task performance were brought to light in the unique working environment of a simulated Mars mission. The hypotheses proposed by this study were (1) the cognitive and affective variables in the COPE model are influenced by the work content, and (2) variation in task performance can be explained by an individual's emotional state and cognitive task load. Both hypotheses were accepted; however, evidence for the second hypothesis was not strong.

In Chapter 4, an experiment was conducted in which the COPE model's variables were used to assess the predictability of performance addressing the second research question. All variables of the COPE model were measured during scenario-based training, performed in a ship simulator at the Royal Netherlands Naval College. Analyses were aimed at the following hypotheses: (1) the variables are related as proposed by the COPE model, and (2) the work content and cognitive and affective factors from the COPE model are predictors for performance and errors. In the third chapter, evidence to support this last hypothesis was meagre. The fourth chapter, however, provided

predictive models, or functions, that could predict both types of performances (performance rated by experts and number of errors made), using the COPE variables.

In the second part of the thesis, a digital support tool based on the COPE model is described. This tool provides biofeedback in combination with real-time predicted performance feedback. The design of this system is described in Chapter 5, and the evaluations are presented in Chapters 6 and 7.

Chapter 6 describes another scenario-based training session that took place in the simulator at the Royal Netherlands Naval College. New participants enrolled in the same stressful scenario used in Chapter 4. A within-subjects experiment was performed; trainees received feedback from the COPE feedback system (physiological, predicted performance and predicted error-chance feedback) during one half of the scenario and no feedback in the other half. The trainees' performance was rated by experts and the number of errors they made was counted using video analysis. The main hypothesis stated that performances improve when feedback is provided. A significant decrease in two types of errors was found, which warrants further research into the effects of individual feedback elements.

The results of Chapter 6 led to the next experiment. In Chapter 7, the effectiveness of the different feedback types and combinations of feedback types from the COPE-FB system are investigated. An experiment in which participants received all possible combinations of feedback while they executed a stressful task was performed. A low-fidelity simulation that resembled a naval setting was created, where the participants needed to extinguish fires that appeared on a ship. This experiment investigated the following hypotheses: (1) providing real-time COPE feedback improves performance and perception of usability, (2) the separate types of real-time COPE feedback improve performance and perception of usability, and (3) combinations of different types of feedback result in an additional positive effect on the improvement of performance and perception of usability. The findings were consistent in concluding that, in general, the providing of feedback positively affects the performance. However, the study was unable to relate this finding to a specific type of feedback or to a specific combination of types of feedback. Still, the results showed a significant user preference for physiological feedback. This preference was lost when an extra error-chance feedback was added to the physiological feedback.

In the general discussion in Chapter 8, the COPE model and its feedback system are reflected upon, including the findings obtained in the various studies. The results of the different experiments suggest improvements for future feedback systems to make them more effective in improving performance under stress. These improvements are also preferred by its users. The results presented in this thesis led to a set of new hypotheses that are discussed in detail in this last chapter.

REFERENCES

Bouchard, S., Bernier, F., Boivin, E., Morin, B., & Robillard, G. (2012). Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. Plos one, 7(4).

Busscher, B., Vlieger, D. d., Ling, Y., & Brinkman, W. P. (2011). Physiological measures and selfreport to evaluate neutral virtual reality worlds. Journal of Cybertherapy & Rehabilitation, 4(1), 15-25.

Cesta, A., Cortellessa, G., & Benedictis, R. D. (2014). Training for crisis decision making - An approach based on plan adaptation. Knowledgde-based systems, 58, 98-112.

Cohn, L. J., Weltman, G., Ratwani, R., Chartrand, D., & McCraty, R. (2010). Stress inoculation through cognitive and biofeedback training. Paper presented at the Interservice/Industry Training, Simulation and Education Conference.

Dörner, D., & Schaub, H. (1994). Errors in planning and decision-making and the nature of human information processing. Applied psychology: an international review, 43(4), 433-453.

Driskell, J. E., & Johnston, J. H. (1998). Stress exposure training. In J. A. Cannon-Browers & E. Salas (Eds.), Making decision under stress: Implications for individual and team training (Vol. 3, pp. 191-218). Washington, DC: American Psychological Association.

Kahnemen, D., Slovic, P., & Tversky, A. (1982). Judgment under Uncertainty: Heuristics and biases. Cambridge, MA: Cambridge University Press.

Keinan, G., Friedland, N., & Ben-Porath, Y. (1987). Decision making under stress: scanning of alternatives under physical threat. Acta psychologica, 64, 219-228.

Kerstholt, J. H. (1994). The effect of time pressure on decision-making behaviour in a dynamic task environment. Acta psychologica 86, 89-104.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. Paper presented at the Proceedings of the Human Factors and Ergonomics Society annual meeting.

Kontogiannis, T., & Kossiavelou, Z. (1999). Stress and team performance: principles and challenges for intelligent decision aids. Safety science, 33, 103-128.

Kowalski-Trakofler, K. M., Vaught, C., & Scharf, T. (2003). Judgment and decision making under stress: an overview for emergency managers. International Journal of Emergency Management, 1(3), 278-289.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? American Psychologist, 59(1), 14.

Ozel, F. (2001). Time pressure and stress as a factor during emergency egress. Safety science, 38, 95-107.

Peeters, M., Van Den Bosch, K., Meyer, J.-J. C., & Neerincx, M. A. (2014). The design and effect of automated directions during scenario-based training. Computers & Education, 70, 173-183.

Raaijmakers, S. F., Steel, F. W., Goede, M. d., Wouwe, N. C. v., Erp, J. B. F. v., & Brouwer, A.-M. (2013). Heart rate variability and skin conductance biofeedback: A triple-blind randomized controlled study. Paper presented at the Humaine Association Conference on Affective Computing and Intelligent Interaction.

Reason, J. (1987). Cognitive aids in process environments: prostheses or tools? International journal of man-machine studies, 27, 463-470.

Salas, E., & Cannon-Browers, J. A. (2001). The science of training: a decade of progress. Annual review of psychology, 52, 471-499.

Salas, E., Driskell, J. E., & Hughes, S. (1996). Introduction: the study of stress and human performance. In J. E. Driskell & E. Salas (Eds.), Stress and Human Performance (pp. 1-45). Hillsdale, NJ: Erlbaum.

Schraagen, J. M., Ormerod, T., Militello, L., & Lipshitz, R. (2012). Naturalistic decision making and macrocognition: Ashgate Publishing, Ltd.

Starcke, K., & Brand, M. (2012). Decision making under stress: a selective review. Neuroscience and Biobehavioral Reviews, 36, 1228-1248.

Zatarain, L. A. (2008). Tanker War: America's First Conflict with Iran, 1987-88: Casemate.

PART 1
MODELLING PERFORMANCE UNDER STRESS

2. THE COGNITIVE PERFORMANCE AND ERROR (COPE) MODEL

2.1 INTRODUCTION

To improve performances under stress, it is necessary to understand how stress influences performances. This chapter proposes a model that explains these influences. Establishing such a model provides an answer to the first research question of this thesis: w*hich aspects of the work content influence the cognitive and affective factors of cognitive task performance?* Existing literature was consulted to answer this question and to create a model that explains performance decline caused by stress. Contrary to other models e.g., (Hart & Staveland, 1988; Salas et al., 1996), the model presented here includes energetical constructs e.g. effort, arousal, activation, fatigue (Robert & Hockey, 1997; Sanders, 1983). These variables can be measured with physiological variables that represent objective stress levels.

In addition, the variables and constructs included in this model have the potential to be measured in real-time. This is needed to fulfil the aim of this thesis to create a real-time feedback system that improves trainees' performances while working in stressful environments. The idea is that variables that influence performance can be measured in real-time and thereby provide real-time feedback. If the variables change according to the environment or the professional working in it, the output of system, i.e. the feedback, also changes according to the current level of the variables.

To create the COgnitive Performance and Error (COPE) model, existing models and theories (Forgas, 1995; Gaillard, 2008; Lazarus, 1999; Mehrabian, 1996; Neerincx, 2003) were consulted. The COPE model demonstrates how work content and cognitive and affective (energetical) factors are influenced by stressful stimuli and how they determine individual performances. Furthermore, this model incorporates new elements: (1) energetical factors that provide an objective measure for individual responsiveness to stress, and (2) the variables can be translated into real-time measures.

2.2 COPE MODEL

A first draft of the COPE model was previously published (Cohen, Brinkman, & Neerincx, 2012). A refined version is presented in figure 1. This model describes the process of performing under stress. It consists of interactions between factors within three dimensions: (1) the work content, (2) the individual's cognition and affect, and (3) the individual's actions.

2.3 MODEL VARIABLES

The following sections describe the variables from the three dimensions presented in

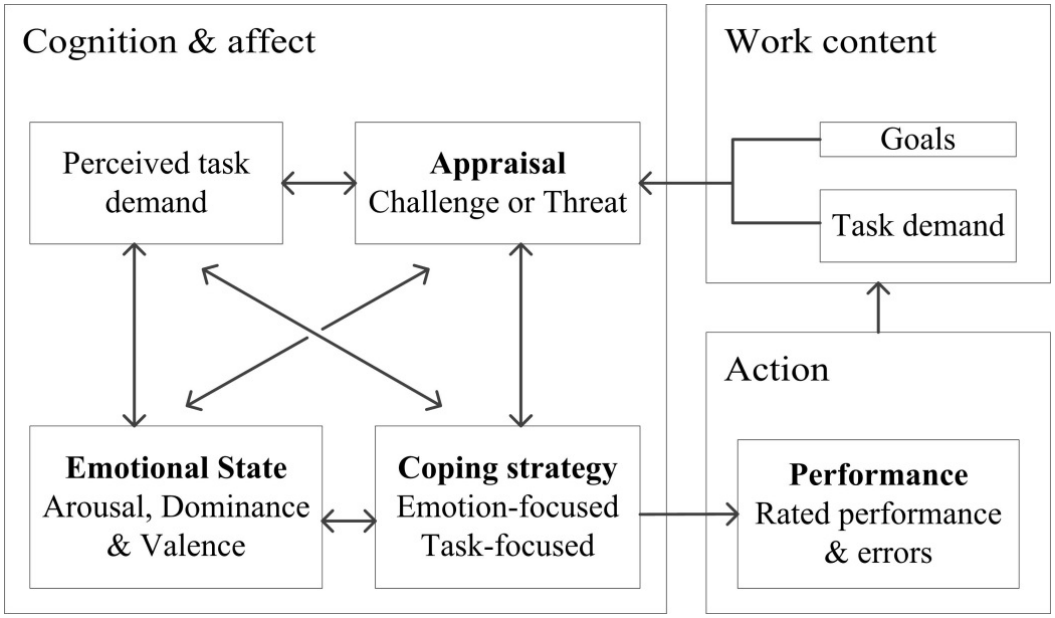


Figure 2.1. The COgnitive Performance and Error (COPE) model

the COPE model (figure 2.1). Relevant literature is cited and the underlying relations between the factors are described. Overall, the COPE model explains how work content influences one's perception of a task and one's affective state. Both of these are factors that influence the ultimate performance.

2.3.1 GOALS

The COPE model distinguishes work content aspects that influence the performance under stress. The specific task *goals* and task demands of the work will characterize the involved cognitive and affective processes (Veltman & Jansen, 2004a; Veltman & Jansen, 2003). *Goals*, often structured in a hierarchical way, drive the performance, but may be appraised differently (e.g., due to its relevance for a higher order goal). More challenging goals improve performance compared to easy goals (Locke, Shaw, Saari, & Latham, 1981). Tasks with different structures or characteristics might also provoke different goals and thereby show differences in task performance.

2.3.2 TASK DEMAND

The *task demands* need to be met to complete the task successfully. This factor is represented by the Cognitive Task Load (CTL) model of Neerincx (2003). This model contains three levels: Time Occupied (TO), Task Set Switches (TSS), and Level of Information Processing (LIP). The TO level is the proportion of time needed to complete the task within the available time frame, and the TSS represents how often tasks were switched. LIP is based on the levels of cognitive processes by Rasmussen (1982) and dual process theories (Evans, 2003). Cognitive processes can also be distinguished on a

continuum from analytical to intuitively. According to Hammond (1988), whether someone's cognitive processing leans more towards one or the other depends on the failure or success of previous judgments, as well as on the task characteristics. As with the cognitive processes, tasks can be placed upon a continuum from 'inducing analytical cognition' to 'inducing intuitive cognition'.

Task demands in the work content dimension of the COPE model are generic for a certain group of individuals. In Section 2.3.4, the perceived task demand is explained, and the difference between the two task demand factors is further illustrated.

2.3.3 APPRAISAL

An individual's reaction to a task is influenced by the meaning, or *appraisal*, that person gives to the task (Lazarus, 1999; Lazarus & Folkman, 1984). When an event is being perceived, the *primary appraisal* will assess the severity of potential danger. When a situation is appraised as possibly harmful, the resources to deal with the danger will be assessed in the secondary appraisal. The *secondary appraisal* leads to the perception of the events that go with the task as either a challenge or a threat. Different individuals may have different appraisals of the same events or stimuli (Anshel, 2000). Some individuals may believe they are able to cope with a stressful task and see it as a challenge, whereas others believe they lack the skills to cope with the task and thus perceive it as a threat. An event with a very high task demand is usually perceived more as threatening than challenging.

Appraisal also influences the current emotional state of the task executor. Perceiving a task as a threat evokes mostly negative feelings, and perceiving a task as a challenge arouses a more positive state. Furthermore, the coping style one uses to handle a task or event is determined by the perception of a task. Threats evoke an emotion-focused coping style, while challenges evoke more task-focused coping styles (Folkman & Lazarus, 1985).

2.3.4 PERCEIVED TASK DEMAND

Perceived task demand is defined by the level of task demand, as described in Section 2.3.2, perceived by the task executor. The perceived task demand and appraisal relate to one another. When a situation is seen as a threat, the perceived task demands will be higher than when a situation is seen as a challenge, and vice versa. The relations between perceived task demands and emotional state and coping are expected to be similar to those between appraisal and emotional state and coping, since both perceived task demand and appraisal represent the individual experience of task perception.

Whether the perceived task demand deviates from the general task demand is determined by several measures, such as experience and emotional state prior to the task. Contrary to level of experience, emotional state is included into the model (Section 1.2.5). In the conclusion (Section 1.3) the lack of a separate experience factor in the

COPE model is explained.

2.3.5 EMOTIONAL STATE

Emotional state is an important factor in decision-making (Mosier & Fischer, 2010). It is defined as a transitory feeling that depends more on the situation than on the person (Larsen & Buss, 2005). According to the PAD-model of Mehrabian (1996), emotional states are characterised by three dimensions: valence (pleasure), arousal (energy), and dominance (control). Mosier and Fischer (2010) note the difference between affect present prior to the task (incidental affect) and affect induced by the task (integral affect). The perception (i.e. appraisal and perceived task demand) of a task affects the (integral) emotional state of the task executor. Demanding and threatening tasks will evoke higher arousal and negative valences, while less demanding tasks that are experienced as a challenge will not evoke arousal or negative valences (Gaillard, 2008).

Incidental affect is expected to influence task perception, since affect influences judgment as indicated by the Affect-Infusion model (Forgas, 1995). Therefore, incidental emotional state affects appraisal, perceived task demand, and coping strategy.

2.3.6 COPING STRATEGY

After work content has affected the cognitive and affective factors, individuals will use *coping strategies* that seems appropriate to the situation (Gaillard, 2007). The literature agrees that there are basic coping strategies used when under stress, such as *emotion-focused coping,* aiming to alter or control emotional distress, and *task-focussed coping,* aiming to alter the task or problem (Endler & Parker, 1994; Endler & Parker, 1990; Lazarus & Folkman, 1984). Task stress also triggers different coping styles in different individuals (Matthews & Campbell, 1998). The coping strategy determines how the task is executed and therefore the level of task performance (Delahaij, 2009).

Note that although coping strategy is an important factor in this model, in the following chapters of this thesis it falls out of the experimental designs. Considering that scientists argue that coping behaviour is too dynamic to be predicted from personality traits (Cohen & Lazarus, 1973; Folkman & Lazarus, 1985), it should be measured with questionnaires. Questionnaires that measure coping strategies, however, were regarded as too long and intrusive to be filled out for each (sub)-task during a scenario-based training. These questionnaires would also conflict with the aim of using real-time measurable factors.

2.3.7 PERFORMANCE

A task executor has a certain competence level that represents what the executor can or cannot accomplish. Depending on cognitive and affective factors, the task performance

can approach or digress from the competence (Matthews, Davies, Westerman, & Stammers, 2008a). Whether or not actions taken by the task executor are appropriate and successful in reaching their goal is called (task) performance. In this thesis, the performance (i.e. appropriateness of the performed actions) is expressed in two ways: (1) the performance levels rated by experts such as trainers, and (2) the errors and error tendencies observed during task execution. Errors occur when planned mental or physical activities fail to reach their expected goal (Reason, 1990), or when the execution of the actions fail.

Both measures of performance are indirectly or directly influenced by cognitive factors and emotional states affected by the work content factors. The performance of a task changes the work content, and thereby alters the goals and task demand. The process of working under stress as described by the COPE model begins again.

2.4 CONCLUSIONS

The COPE model proposed in this chapter describes how work content factors influence cognitive and affective factors of cognitive task performance, and thereby indirectly influence task performance itself. The variables and their (expected) relationships have been described. Although elements, such as experience and knowledge, are important in these processes (Klein, 1993; Noble, 1998), it is believed that the cognitive and affective variables in the COPE model reflect the level of experience. For example, an expert will more often than a novice appraise an event or task as a challenge rather than a threat. Nevertheless, the experiments presented in this dissertation focus on participants with the same level of experience and prior knowledge (i.e., we will model the relationships between these variables for rather coherent groups of participants). The experiments will also exclude the factor of coping strategy. Questionnaires for coping strategy consist of multiple scales and are therefore not desirable for the purpose of this thesis. The variables included in the COPE model all have the potential to be measured either through automatic physiological measures, or through short, one scale questionnaires. If such measures become available for coping strategy, it is advised to include this variable in the COPE model.

The next two chapters describe the next step, namely the validating of the COPE model in experimental settings. Chapter 3 validates parts of the model with a dataset collected during a simulated Mars-mission, and Chapter 4 validates the COPE model with data collected during a simulated Naval mission.

REFERENCES

Anshel, M. H. (2000). A conceptual model and implications for coping with stressful events in police work. Criminal Justice and Behavior, 27(3), 375-400.

Cohen, F., & Lazarus, R. S. (1973). Active coping processes, coping dispositions, and recovery from surgery. Psychosomatic Medicine, 35(5), 375-389.

Cohen, I., Brinkman, W. P., & Neerincx, M. A. (2012). Assembling a synthetic emotion mediator for quick decision making during acute stress. Paper presented at the Proceedings of the 2012 European Conference on Cognitive Ergonomics, Edinburgh.

Delahaij, R. (2009). Coping under acute stress: the role of person characteristics. Kon. Broese & Peereboom, Breda.

Endler, N. S., & Parker, J. D. (1994). Assessment of multidimensional coping: Task, emotion, and avoidance strategies. Psychological assessment, 6(1), 50.

Endler, N. S., & Parker, J. D. A. (1990). Multidimensional assessment of coping: a critical evaluation. Personality processes and individual differences, 58(5), 844-854.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. Trends in Cognitive Sciences, 7(10), 454-459.

Folkman, S., & Lazarus, R. S. (1985). If it changes it must be a process: study of emotion and coping during three stages of a college examination. Journal of personality and social psychology, 48(1), 150.

Forgas, J. P. (1995). Mood and judgement: the affect infusion model (AIM). Psychological bulletin, 117(1), 39-66.

Gaillard, A. (2007). Stress productiviteit en gezondheid (Vol. 3). Amsterdam: Holland Graphics.

Gaillard, A. W. (2008). Concentration, stress and performance. Performance under stress, 59-75.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in psychology, 52, 139-183.

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making Decision making in action: Models and Methods: Ablex Publishing Corporation.

Larsen, R. J., & Buss, D. M. (2005). Personality Psychology: Domains of Knowledge about Human Nature (2nd ed.). New York: McGraw-Hill Higher Education.

Lazarus, R. S. (1999). Stress and emotion: a new synthesis. New York: Springer Publishing Company, Inc.

Lazarus, R. S., & Folkman, S. (1984). Stress, Appraisal, and Coping. New York: Springer Publishing Company, Inc.

Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969–1980. Psychological bulletin, 90(1), 125.

Matthews, G., & Campbell, S. E. (1998). Task-induced stress and individual differences in coping. Paper presented at the Humand factors and ergonomics society annual meeting.

Matthews, G., Davies, D. R., Westerman, S. J., & Stammers, R. B. (2008). Human Performance: cognition, stress and individual differences (3 ed.). New York: Psychology Press.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology, 14(4), 261-292.

Mosier, K. L., & Fischer, U. (2010). The role of affect in naturalistic decision making. Journal of cognitive engineering and decision making, 4(3), 240-255.

Neerincx, M. A. (2003). Cognitive task load design: model, methods and examples. In E. Hollnagel (Ed.), Handbook of Cognitive Task Design (pp. 283-305). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Noble, D. (1998). Distributed situation assessment. Paper presented at the Proc. FUSION.

Rasmussen, J. (1982). Human Errors: A taxonomy for describing human malfunction in industrial installations. Journal of Occupational Accidents, 4, 311-333.

Reason, J. (1990). Human error: Cambridge university press.

Robert, G., & Hockey, J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. Biological Psychology, 45(1), 73-93.

Salas, E., Driskell, J. E., & Hughes, S. (1996). Introduction: the study of stress and human performance. In J. E. Driskell & E. Salas (Eds.), Stress and Human Performance (pp. 1-45). Hillsdale, NJ: Erlbaum.

Sanders, A. (1983). Towards a model of stress and human performance. Acta psychologica, 53(1), 61-97.

Veltman, H. J., & Jansen, C. (2004). The adaptive operator. Human performance, situation awareness, and automation: Current research and trends, 2, 7-10.

Veltman, J. A., & Jansen, C. (2003). Differentiation of mental effort measures: consequences for adaptive automation. In G.R.J.Hockey, A. W. K. Gaillard & O. Burov (Eds.), Operator Functional State: The Assessment and Prediction of Human Performance Degradation in Complex Tasks (pp. 249-259). Amsterdam: IOS Press.

3. Work content influences on cognitive task load, emotional state and performance during a simulated 520-days' Mars mission

Abstract

In high-risk domains such as human space flight, cognitive performances can be negatively affected by emotional responses to events and conditions in their working environment (e.g., isolation and health incidents). The COgnitive Performance and Error (COPE) model distinguishes effects of work content on cognitive task load and emotional state, and their effect on the professional's performance. This chapter examines the relationships between these variables for a simulated Mars-mission. Six volunteers (well- educated and -motivated men) were isolated for 520 days in a simulated spacecraft in which they had to execute a (virtual) mission to Mars. As part of this mission, every other week, several computer tasks were performed. These tasks consisted of a negotiation game, a chat-based learning activity and an entertainment game. Before and after these tasks, and after post-task questionnaires, the participants rated their emotional state consisting of arousal, valence and dominance, and their cognitive task load consisting of level of information processing, time occupied and task-set switches. Results revealed significant differences between cognitive task load and emotional state levels when work content varied. Significant regression models were also found that could explain variation in task performance. These findings contribute to the validation of the COPE model and suggest that differences in appraisals for tasks may bring about different emotional states and task performances.

Keywords: emotional state, cognitive task load, performance, stress, human space flight.

Chapter submitted as:

Cohen, I., den Braber, N., Smets, N.J.J.M., van Diggelen, J., Brinkman, W.P. & Neerincx, M.A. (2015) Work content influences on cognitive task load, emotional state and task performance during a simulated, 520 day's Mars mission. *Computers in human behaviour.*

3.1 INTRODUCTION

Different professionals, such as police officers, military personnel, pilots and astronauts, occasionally enter high-risk situations, in which the risk for harm, information uncertainty and time pressure evoke stress in the professionals involved (Driskell & Johnston, 1998). Their job is to remain focused and perform well in these situations. Extreme levels of stress, however, can affect cognitive performances in negative ways and consequently deteriorate performances (Keinan et al., 1987; Ozel, 2001; Starcke & Brand, 2012).

Insight into human and work content factors that determine cognitive task performance in these situations are useful for finding ways to counteract the performance decline. When the influences of these factors are known, the focus of support can be placed where the help is needed. It might also allow for better anticipation for such situations (e.g., an improved human resource deployment). By monitoring the human and content variables that affect task performance, content-sensitive and personalized task support can be provided.

Based on a literature study and domain analyses, Cohen et al. (2012) proposed the COgnitive Performance and Error (COPE) model as a general foundation for task support in high-risk domains. In several empirical studies, this model was refined, "parameterized" and evaluated for different application domains. This chapter studies the influences from different work contents on core variables of the COPE model (i.e. cognitive task load and emotional state) and the prediction of task performance based on these variables. The analysis centres around a unique experiment on human space flight: the Mars500 program[1] (i.e., a simulated complete, 520-days' Mars mission of a group of six astronauts).

In the Mission Execution Crew Assistant (MECA) project, as part of the Mars500 program, the astronauts performed a set of tasks every two weeks under the stressful conditions of a long-duration mission. This experiment was set-up to refine and test the MECA requirements baseline for electronic partners (ePartners) that enhance astronaut-automation groups' performance and resilience (Neerincx, 2011; Neerincx et al., 2008; Smets, Cohen, Neerincx, Brinkman, & Diggelen, 2012). MECA is developing personal ePartners that regularly monitor crew-members *cognitive task load* and *emotional states* during individual and joint task performances over all mission phases (Neerincx et al., 2008). This monitoring is a joint crew-ePartner activity and the basis of envisioned ePartner support functions that should help to better cope with the social, cognitive and affective burdens arising in such environments (Diggelen & Neerincx, 2010; Gorbunov, Barakova, Ahn, & Rauterberg, 2011; Hennes, Tuyls, Neerincx, & Rauterberg, 2009). The COPE-model from the previous chapter (Chapter 2) might provide a basis. Before the study is presented, this chapter will discuss different work content factors that affect performances during long-term isolation missions.

[1] www.esa.int/Our_Activities/Human_Spaceflight/Mars500

3.1.1 *WORK CONTENT*

In Chapter 2, the COPE-model was presented. COPE consists of three components: work content, cognitive and affective factors, and the actions. The work content is divided over goals, and task demand.

Goals that need to be achieved by the task performer are often hierarchical structured. Completing an overall training to learn a certain skill is a higher level goal. Such goals will be accomplished by achieving different lower level goals, or sub-goals, such as learning different components of the skill. For some lower level goals, the link to the higher level goals will not be as obvious as for other lower level goals, e.g., the mapping of the work goals on the computer tasks is not straightforward (cf. (Kieras & Polson, 1985; Sutcliffe, Ryan, Doubleday, & Springett, 2000). This mismatch between the different hierarchical goals will be visible in the perception of the work content. In other words, the appraisal of the work content is dependent on the fit between the lower level and higher level goals.

The simulated Mars mission described in this chapter also contains goals from different levels. A few higher level goals will be present during different phases of the mission. Long-term missions to the ISS and MIR space stations have been divided according to a stage model by Manzey (2004) and Gushin, Kholin, and Ivanovsky (1993). These missions last between 4 to 6 months and every stage has its own psychological stressors. In the first phase, that last approximately 4 to 6 weeks, crewmembers mainly focus on adaptation to the physiological changes. Stress and performance problems in this phase are induced by these physical adaptations. Full adaptation to the new conditions is reached in the second phase that is followed by the most difficult third phase, where psychological problems are likely to occur. This third phase starts after approximately 6 to 12 weeks in space. Severe stressors in this phase are: monotony and (social) boredom, isolation from family and friends, and the omnipresent contact with the other crewmembers. The fourth phase starts shortly before the end of the mission. It evokes euphoria but also concerns as to ending and completing the mission. Within these different stages during a long-term mission, different higher level goals can play a role. In addition to the goal of exploring the Mars surface, the Mars 500 mission distinguishes four phases with corresponding (higher level) goals: (1) adapt towards the (new) space environment, (2) establish efficient work procedures or routines, (3) prepare for the Mars landing, and (4) return to home (Earth). The different computer tasks that need to be performed during the different phase of the missions have lower level goals that, ideally, would contribute to such higher level goals (i.e., support the adaptation, the routine development, the landing preparation and the return).

3.1.2 *RESEARCH QUESTIONS*

In an effort to gather knowledge on psychological effects of a Mars mission on its

crewmembers, the European Space Agency (ESA) and the Institute of Biomedical Problems (IBP) carried out the Mars500 project. This project simulated a Mars mission in its full length of 520 days here on Earth including the isolation factors and the lack of contact with Earth. Obviously, the absence of gravity was not reproduced but the unique settings of a Mars mission simulation brought its own unique stress-factors. During the experimental sessions in the Mars 500 project, emotional state (ES) and cognitive task load (CTL) were measured while the participants executed different tasks. In the COPE-model, ES and subjective CTL (i.e., perceived task demand) influence each other, and determine what the results will be of the task that is being performed. This leads to the following two hypotheses investigated in this study:

- The cognitive and affective variables in the COPE model are influenced by work content:
    a. The fit of a task goal with a higher level goal, reflects in the levels of the ES and CTL levels.
    b. Different overall mission phases evoke different levels of the ES and CTL levels.
- Variation in task performance can be explained by individuals' emotional state and perceived task load.

It was expected that different tasks with different task goals, evoke different levels of ES and CTL. The different task goals can either fit with the higher level goals. If this is the case, valence, arousal and dominance are expected to increase compared to emotional state levels of tasks with unfitting goals. The same can be expected for cognitive task load. If a task is appropriate for reaching a higher-level goal, CTL will increase.

The phase of the mission was also expected to influence these variables since different phases are related to higher-level goals, or not. Phases during a Mars mission are, however, of different nature than mission phases during MIR and ISS missions (Gushin et al., 1993; Manzey, 2004). For one, a Mars mission lasts 520 days instead of 2 or 3 months. Therefore, the influences of mission phases on ES and CTL levels are expected to differ from those in previous studies with MIS and ISS crewmembers. Euphoria caused by returning home, is present in ISS and MIR mission since the return home takes a few days. Returning home from Mars takes approximately 6 months and a euphoric feeling based on a return mission is not expected in the last phase.

- Arousal is expected to decline during the entire mission since crewmembers get adapted to the situation. At the end of the mission, they are not as excited as at the beginning.
- Valence is expected to be quite stable over the mission. Except the period around the Mars landing, where a high valence is expected. Since the mission phases in the Mars500 project are quite long, this effect might not be strong enough to be visible in one phase compared to other phases.
- Dominance is expected to act in the same manner as valence.
- Cognitive task load is expect to decrease over the course of the mission. The tasks

have been performed for a while and do not cause as much CTL as before. The perceived task demand might increase at the end of the mission when crewmembers are more fatigued.

According to the COPE model, the cognitive and affective factors influence performances. It was therefore expected that variation in the two factors could be associated with observed performance variation.

3.2 METHODS

The study had a longitudinal correlational design. Over a period of 520 days, multiple observations were made with regard to emotional state, cognitive task load and task performance on the same set of tasks.

3.2.1 PARTICIPANTS

A total of six male participants with a mean age of 32.3 (minimum 27, maximum 38 years) were selected for the Mars 500 project. The selection procedure required male volunteers between 25 and 50 years of age with a higher education degree. The participants were divided into two groups of three participants. These were also the groups in which the tasks were performed. For practical reasons, one group consisted of the English speaking participants and one group consisted of the Russian speaking participants.

3.2.2 WORK CONTENT (TASKS)

Every other week, a session started for half an hour. In every session three tasks were executed: a learning activity, called Collaborative Trainer (COLT); a negotiation game, named Colored Trails (CT); and an entertainment game, called Lunar Lander (LL). COLT and CT are multi-user (group) tasks, whereas LL is a single-user game. The different tasks are explained in the next sections.

COLORED TRAILS (CT)

Colored Trails is a negotiation game with competitive elements for two or more users. This game is developed as a research test-bed for investigating decision-making in groups and proposed as a tool for assessing group-members' relationships and (a-)social behaviours towards each other (Gorbunov et al., 2011). The three group members played the game on a rectangular board with coloured squares (see figure 3.1a). Group members had their own piece on the board, which they could move with a

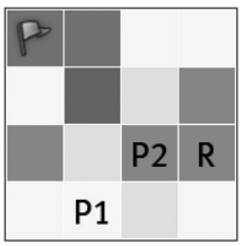
coloured chip. The general goal was to position pieces as close as possible to the flag. All players saw the board and the chips possessed by the other players, which made it possible to propose chip exchange. A player who received propositions could either except one or decline all. According to a specific scoring function, each player could earn points with its moves. The game-time was around 10 minutes (for a more detailed description of the game and analyses of the group-members' CT-performances and relationships, see Gorbunov et al. (2011)).

LUNAR LANDER (LL)

Lunar Lander is an entertaining game, played individually. This version was a Java-version of the original 1979's Lunar Lander video game from Atari. A player had to land a space-ship safely on the surface of the moon as many times as possible without crashing (see figure 3.1b). The difficulty level increased successively. Pressing the arrow buttons altered the space-ships direction and the spacebar accelerated the space-ship forward.

COLLABORATIVE TRAINER (COLT)

Collaborative Trainer was a learning task for three persons, one teacher (instructor) and two students. The students' goal was to learn procedures for the usage, maintenance and damage-control of systems. The teacher had to provide the assignments and to guide student's learning processes (i.e., pointing to the relevant learning material and giving hints when needed) . This way, COLT combines computer-based learning and collaborative learning techniques. The teacher sent instructions via chat to the students, who then executed the specific task. For each assignment, the teacher had background information available on his dashboard to supervise, help and advice the students while they were learning. COLT was used to learn the relevant procedures of two different systems: Cardiopres and Watertank. Cardiopress is a real payload for physiological measurements in space stations (ECG, breathing, skin conductance, blood pressure), and COLT contained all "official" procedures and background (multimedia) information for its usage, maintenance and Fault Detection,



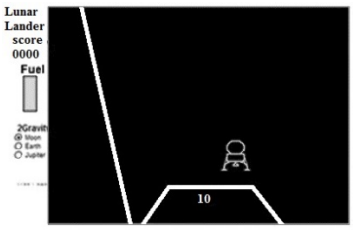


Figure 3.1a. Colored Trails game board 3.1b. a screenshot of Lunar Lander

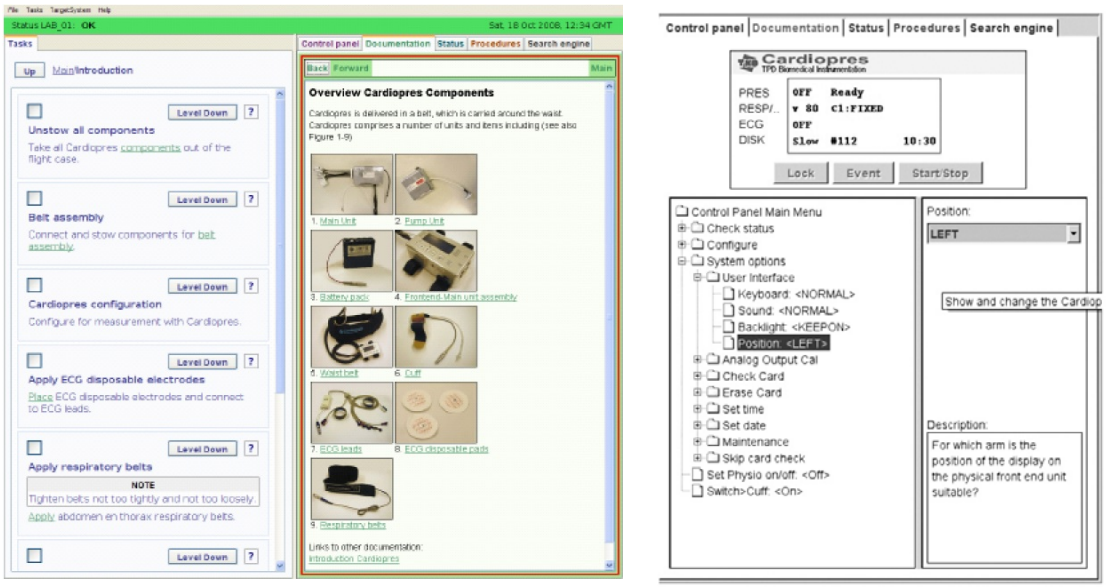


Figure 3.2. Screenshots of the COLlaborative trainer task, Cardiopres version.

Isolation and Recovery (FDIR) procedures (see figure 3.2). The Watertank system was a simplified simulator of a hypothetical water provision system, for which COLT provided some derived procedures for usage and fault recovery. The Watertank scenario was always played with the same teacher, whereas for the Cardiopress the teacher role rotated among the three group members (for more details on COLT, see Smets et al. (2012)).

3.2.3 Work content: Phases

The whole simulated Mars mission lasted for 520 days and was divided into four different phases corresponding the stage model described by Manzey (2004) and Gushin et al. (1993). The simulated Mars landing divided the mission into two halve. Both halves were divided equally, resulting in four phases. The first phase (session 1 to 9, week 1 to week 18) and the second phase (session 10 to 19, week 19 to 38) were before the Mars landing, and phase three (session 20 to 29, week 39 to week 58) and four (session 30 to 38, week 59) were after the Mars landing.

3.2.4 Measures

Several variables were collected to measure the abstract constructs of the COPE model: emotional state, cognitive task load, and task performance.

Emotional State (ES) & Cognitive Task Load (CTL)

A common way of measuring emotional state is by using the Self-Assessment Manikins

from Bradley and Lang (1994). This questionnaire consists of three 5-point-likert scales on valence, arousal and dominance. While valence is a scale that indicates the pleasantness of stimuli experienced by an individual, the arousal scale indicates the activation level. Dominance represents the level of control an individual feels over certain stimuli or situations. Every point on the scales was represented by a small icon as shown in figure 3.3. The three levels of cognitive task load were also measured on a 5-point scale. This questionnaire is shown in figure 3.3 as well.

Next to the rated ES and CTL scores, the difference between two of these scores were also used in the analyses of this study. For example, valence was measured before a task, and after a task. The Δ valence was used as an indication of valence change. In Section 2.7 the different measurement moments of ES and CTL are explained, and Section 3.1 explains the Δ's variables in more detail.

PERFORMANCE

All three tasks aim at achieving individual task or learning goals; the performance scores were determined in different ways. During Lunar Lander, points were received for every successful landing. The score that could be achieved for a landing on a particular spot was visible underneath the surface of that spot as shown in figure 3.1b.

For Colored Trails the score was calculated as follows; reaching the goal location would deliver 125 points. For not reaching the goal, 25 penalty points were subtracted for every square between the goal and the player's position. In addition, for every chip the player had not used, he received 10 extra points.

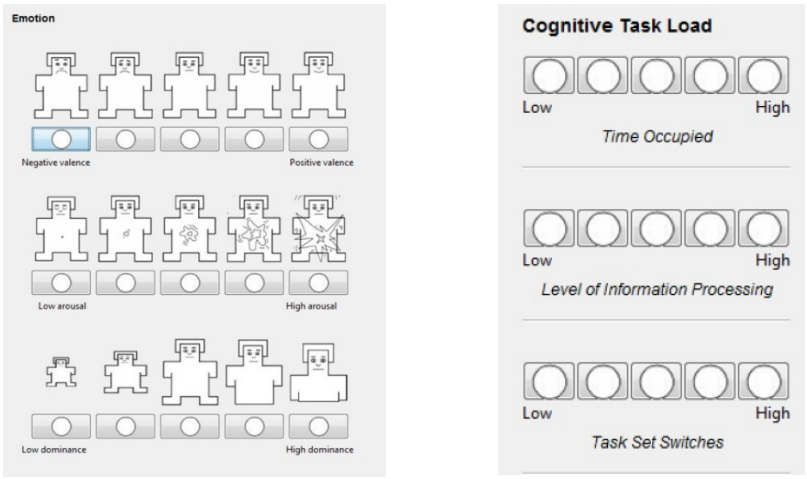


Figure 3.3. The Emotional State questionnaire on the left, and the Cognitive Task load questionnaire on the right.

After performing the COLT tasks, a task questionnaire was filled in, in which the retention of the knowledge gathered during the task was examined. It asks questions about facts and procedures. This was followed by a questionnaire asking the teacher to score the students and himself, and asking him to ask students to rate his teacher performance. Scores were on a 5-point scale: from 1 (poor) to 5 (good). Students also received a similar questionnaire, asking them to rate their own performance, and asking another participant to score their performance.

3.2.5 EXPERIMENTAL DESIGN

The experiment had a repeated measures design. Over a period of 520 days, every other week, multiple observations were made with regard to emotional state, cognitive task load and task performance during the execution of three computer-based tasks.

3.2.6 PROCEDURE

Every two weeks the groups performed a session for half an hour, consisting of all three games: Lunar Lander, Colored Trails and one of the COLlaborative Trainer tasks. First the participants logged on to the system and a timer, a chat client, and an overview screen started. The timer and chat client were on during the whole experiment. Next, a game performance screen was shown, followed by the first game or evaluation task of that session. All three games followed an almost similar procedure (figure 3.4). The task starts with an emotional state questionnaire (time = T0), followed by the tasks. After the task was completed an emotional state and a cognitive task load questionnaire followed (time = T1). For the Lunar Lander and Colored Trails task, the procedure stops there. The COLT sequence continued with an examination part, followed by a teacher/student questionnaire and a second emotional state and cognitive task load questionnaire (time= T2).

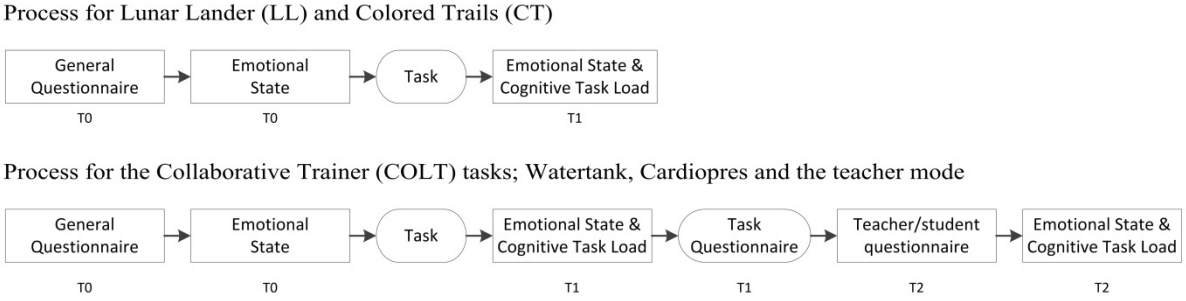


Figure 3.4. procedure for the different tasks. Top picture shows Lunar Lander and Colored Trails procedure, bottom picture shows the procedure for the COLT tasks.

3.3 RESULTS

All analyses were executed in R Studio and an alpha level of .05 was used for all statistical tests. Before the analyses were conducted, the data needed preparation.

3.3.1 DATA PREPARATION

Reliability analyses in table 3.1 showed a high level of consistency between the three cognitive task load measures (LIP, TOC, and TSS). The three separate levels were replaced for a single aggregated mean score for cognitive task load that was used as a predictor in the regression analyses and as dependent variables in the ANOVAs. An extra variable was created with the difference between ES measurement at T1 and T0 ($value_{T1} - value_{T0} = value_{\Delta T}$). Another extra variable was created to indicate the phase of the mission. Phase 1 lasted from session 1 up to and including session 9, phase 2 included session 10 up to 19, phase 3 consisted of sessions 20 up to 29 and phase 4 included sessions 30 up to 38.

The small group of participants in this study might be "interesting in themselves" and create a "sample that exhausts the population" which are indications for fixed effects (Gelman, 2005). When this is the case, the participants can be treated as fixed effects (Mirman, Dixon, & Magnuson, 2008). In all the ANOVA's and multiple linear regressions described in the result section, participants were treated as a fixed effect by adding a categorical participant variables into the models.

3.3.2 WORK CONTENT - TASK DIFFERENCES

To test the first hypothesis, a series of one-way ANOVA's were conducted to examine if the cognitive task load and emotional state variables varied when different tasks were

Table 3.1 Cronbach's alpha values for three levels of Cognitive Task Load; LIP, TOC and TSS.

	Cronbach's Alpha	
Task	Standardized CTL	Unstandardized CTL
COLT T1	0.92	0.94
COLT T2	0.96	0.95
Lunar Lander T1	0.93	0.98
Colored Trails T1	0.87	0.93

executed. For all these ANOVAs tasks was the independent variable with 5 levels, i.e. COLlaborative Trainer (3 versions), Colored Trails, and Lunar Lander. The dependent variables were the aggregate cognitive task load level, and the emotional state levels, i.e. arousal, dominance, and valence, measured at T1 and T2, and ΔT1. The results of the ANOVAs are presented in Table 3.2 and show a significant effect on cognitive task load at T1, on valence at T2, and on valence at ΔT1 and on arousal and dominance at ΔT2. Additional Tukey's post-hoc tests showed between which tasks the differences were found. The bar graphs in figure 3.5 a-e show all the significant differences.

Table 3.2. Results of the ANOVAs showing effects of task on CTL, arousal, dominance and valence at different measurement moments.

	df1	*df2*	Sum of Squares	F	*p*
CTL1	4	512	11.68	6.93	***<0.001
CTL2	2	136	2.23	2.59	0.079
T0 Arousal	4	515	3.19	2.05	0.086
T0 Dominance	4	515	0.18	0.19	0.943
T0 Valence	4	515	0.30	0.54	0.706
T1 Arousal	4	513	0.49	0.43	0.787
T1 Dominance	4	513	1.51	1.64	0.162
T1 Valence	4	513	3.03	1.54	0.190
ΔT1 Arousal	4	516	0.32	0.36	0.837
ΔT1 Dominance	4	516	1.42	2.32	0.056
ΔT1 Valence	4	516	3.13	3.14	* 0.014
T2 Arousal	2	136	0.002	0.003	0.997
T2 Dominance	2	136	0.69	1.49	0.229
T2 Valence	2	136	3.88	4.23	* 0.017
ΔT2 Arousal	2	145	66.42	5.42	** 0.005
ΔT2 Dominance	2	145	71.94	3.35	* 0.038
ΔT2 Valence	2	145	133.30	2.38	0.097

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, . $p<0.1$

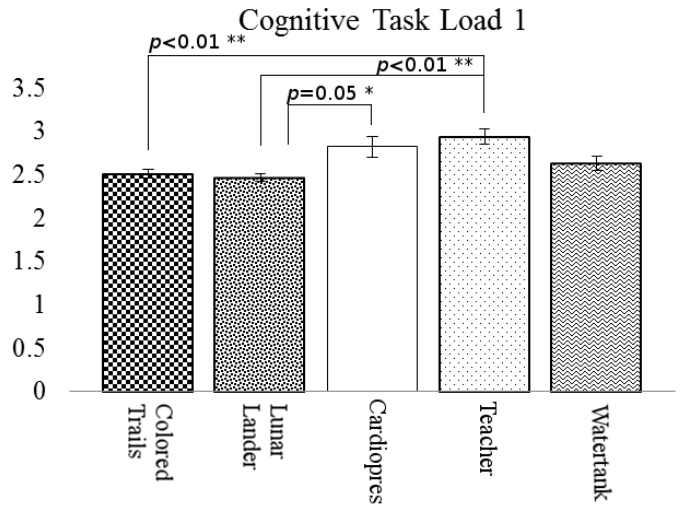


Figure 3.5 a. Tukey's posthoc results for CTL1 differences between tasks.

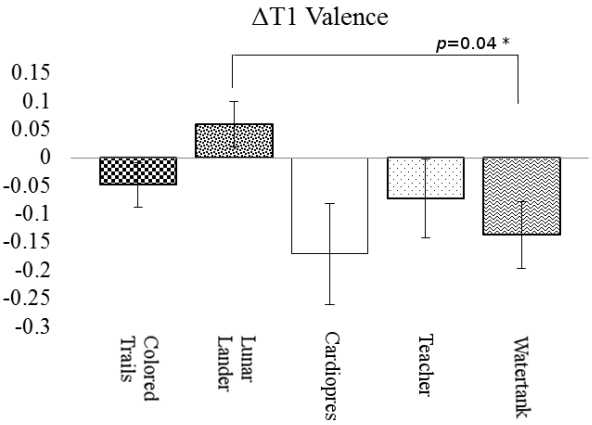


Figure 3.5 b. Tukey's posthoc results for ΔT1 Valence differences between tasks

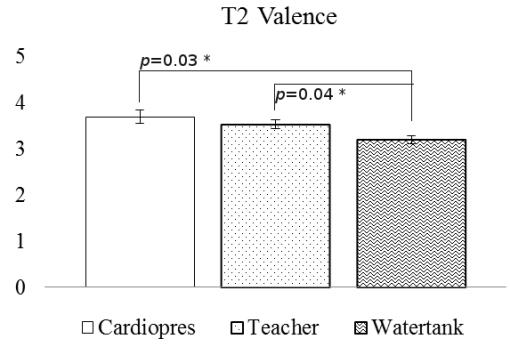


Figure 3.5 c. Tukey's posthoc results for T2 Valence differences between tasks.

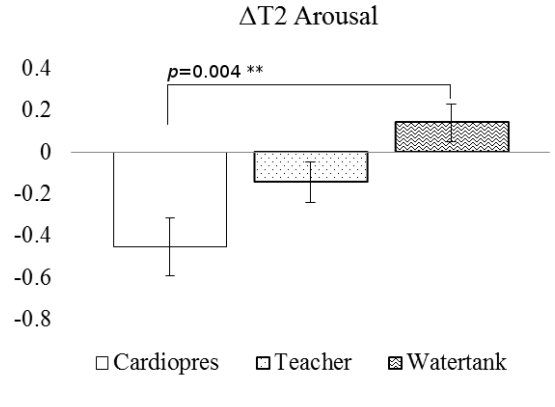


Figure 3.5 d. Tukey's posthoc results for ΔT2 Arousal differences between tasks.

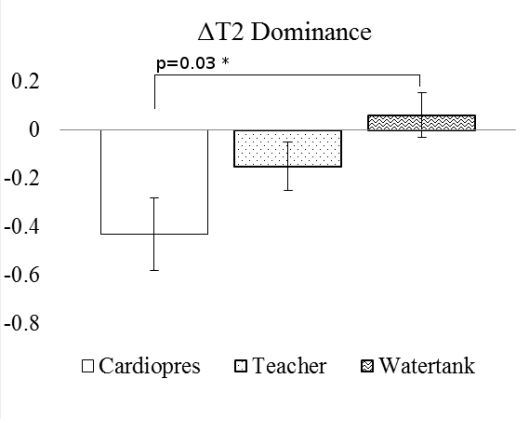


Figure 3.5 e. Tukey's posthoc results for ΔT2 Dominance differences between tasks.

3.3.3 WORK CONTENT - PHASE DIFFERENCES

Further investigations into the first hypothesis looked at differences in CTL and ES variables depending on the phase of the mission in which these values were measured. Details of these ANOVA's are displayed in table 3.3. Differences between phases were found for Cognitive Task Load at T1, arousal at T0, T1 and T2, and valence at T0 and ΔT1. For the dominance level of Emotional State no differences between phases were found suggesting that this level did not vary during the simulated Mars mission. Differences at T0 are most interesting, as the ES levels have not yet been affected by executing tasks.

Tukey's posthoc tests were conducted to examine difference between phases in more detail. Three expected increases or decreases were found. Arousal decreases from phase 2 to phase 3 at T0 and T2 (figure 3.6a and 3.6e). Valence at T0 increased between phase 1 and 2 (figure 3.6b). More differences were found between the non-adjacent phases shown in the bar graphs presented in figure 3.6.

Table 3.3. Results of the ANOVAs showing effects of the different phases on CTL, arousal, dominance and valence at different measurement moments.

		df1	*df2*	Sum of Squares	F	*p*
	CTL1	*3*	*513*	*3.99*	3.06	* 0.03
	CTL2	3	135	0.72	0.54	0.65
T0	Arousal	3	516	4.84	7.31	*** <0.001
T0	Dominance	3	516	0.71	1.76	0.16
T0	Valence	3	516	4.59	3.98	** 0.01
T1	Arousal	3	514	4.87	5.88	*** <0.001
T1	Dominance	3	514	1.31	1.91	0.13
T1	Valence	3	514	1.89	1.23	0.28
ΔT1	Arousal	3	517	0.11	0.17	0.92
ΔT1	Dominance	3	517	0.36	0.78	0.51
ΔT1	Valence	3	517	2.22	2.96	* 0.03
T2	Arousal	3	135	3.48	4.49	** 0.005
T2	Dominance	3	135	0.86	1.24	0.30
T2	Valence	3	135	2.12	1.49	0.22
ΔT2	Arousal	3	144	0.57	0.39	0.76
ΔT2	Dominance	3	144	0.60	0.39	0.76
ΔT2	Valence	3	144	1.85	0.65	0.58

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, . $p<0.1$

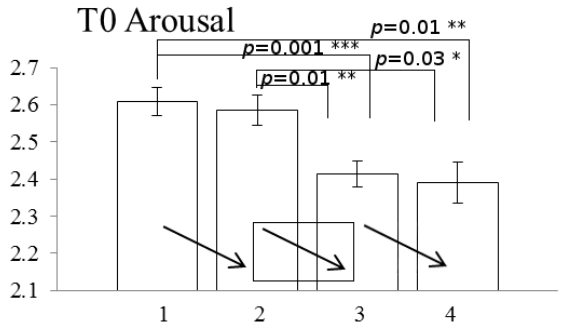


Figure 3.6 a. Tukey posthoc results for T0 Arousal differences between the simulation phases.

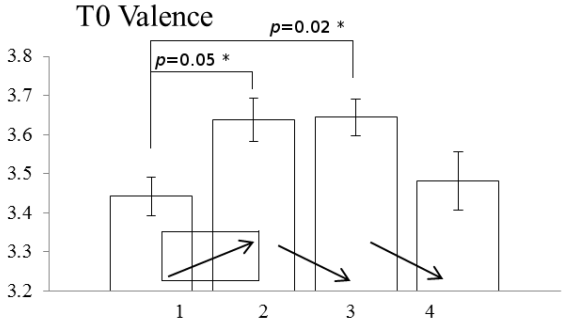


Figure 3.6 b. Tukey posthoc results for T0 Valence differences between the simulation phases.

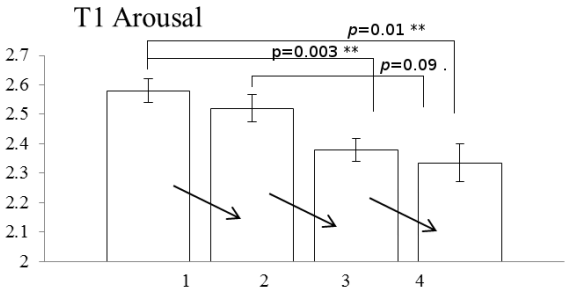


Figure 3.6 c. Tukey posthoc results for T1 Arousal differences between the simulation phases.

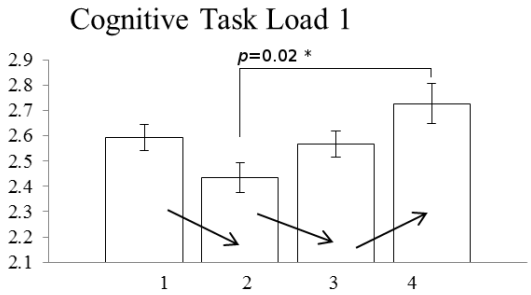


Figure 3.6 d. Tukey posthoc results for CTL1 differences between the simulation phases.

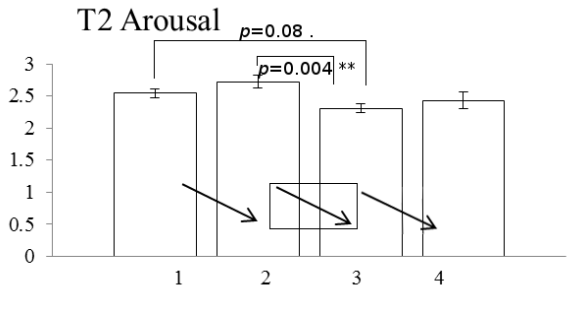


Figure 3.6 e. Tukey posthoc results for T2 Arousal differences between simulation phases.

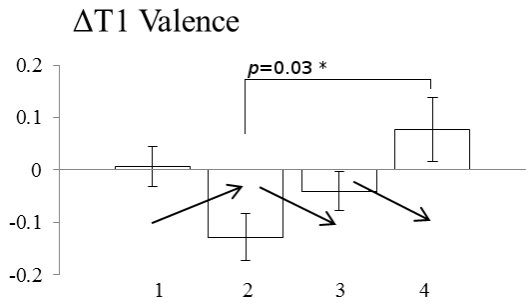


Figure 3.6 f. Tukey posthoc results for ΔT1 Valence difference between simulation phases.

3.3.4 Explaining Task Performance Variation

To find predictors for performance score, regression analyses were conducted. The emotional state and cognitive task load variables showed differences between tasks, therefore, different regression analyses were conducted per task. Table 3.4 shows the results of regression analyses for Lunar Lander task and Colored Trails task. The model

Table 3.4. Regression analyses for the Lunar Lander task and the Colored Trails task. Task performance is the dependent variable and the ES and CTL levels are the independent variables.

Lunar Lander	Estimate	Std. Error	*t*	*p*
Intercept	-137.53	133.44	-1.03	0.304
T0 Arousal	-60.95	45.82	-1.33	0.185
T0 Dominance	142.09	59.68	2.38	*0.018
T0 Valence	3.33	30.64	0.11	0.914
CTL1	9.64	22.74	0.42	0.672
ΔT1 Arousal	3.89	37.90	0.10	0.918
ΔT1 Dominance	87.23	42.84	2.04	*0.043
ΔT1 Valence	55.86	36.73	1.52	0.130
Colored Trails				
Intercept	78.22	29.39	2.66	**0.009
T0 Arousal	-2.04	6.59	-0.31	0.757
T0 Dominance	10.02	9.01	1.11	0.268
T0 Valence	4.08	4.36	0.93	0.351
CTL1	-4.36	4.25	-1.03	0.307
ΔT1 Arousal	-14.45	7.45	-1.94	. 0.054
ΔT1 Dominance	11.42	9.95	1.15	0.252
ΔT1 Valence	2.96	5.81	0.51	0.611

*** p<0.001, ** p<0.01, * p<0.05, . p<0.1

for Lunar Lander was able to account for 40.5% of the performance variance, $F(12, 175) = 11.59$, $p<0.001$. The dominance levels before the task, and the change in dominance were significant predictors and both had a positive association with the variation in task performance. No significant model was found for the Colored Trails task, $F(12, 172) = 0.99$, $p=0.46$.

The three COLT versions resulted in the three models shown in table 3.5. For these tasks, an additional difference between T2 and T1 (valueT2 – valueT1 = valueΔT2) was entered as an indicator for emotional change associated with the examination part. None of the variance in task performance during the Cardiopres task could be accounted for by the variables in the model ($F(13, 15)=1.33$, $p = 0.30$).

For the teacher task, 64.4% of the variance in task performance could be attributed to the model ($F(16, 29)=6.08$, $p<0.001$). The model intercept and cognitive task load at T2 had a positive relation with task performance for this task. CTL at T1 also showed a trend.

Variance in task performance during the Watertank task could be accounted for by the variable in the model for 74.8% ($F(15, 50) = 13.87$, $p <0.001$). The model intercept, arousal at T0 and both valence at ΔT1 and ΔT2 had a positive influences on task performance while dominance at T0 and at ΔT1, had a negative influence on

Watertank's task performance. The dominance at ΔT2 was not significant but showed a trend with a negative coefficient.

A drawback of including a relatively large number of predictors in the analysis is the increased chance of making a Type I error. A more conservative view therefore would be to lower the alpha level considering the number of ES and CTL predictors entered into the model (Mundfrom, Jamis, Schaffer, Piccone, & Roozeboom). For example for the COLT task with 11 ES and CTL predictors, a Bonferroni correction would lower the alpha threshold to .05/11 = 0.0045. For the analysis on the Watertank task, the *p*-value for ΔT1 Valence is still below this alpha level.

Table 3.5. Regression analyses for the three COLT tasks: Cardiopres, teacher and Watertank. Task performance is the dependent variable and the ES and CTL levels are the independent variables.

Cardiopress	Estimate	Std. Error	*t* value	Pr(>|*t*|)
Intercept	4.40	2.10	2.09	. 0.054
T0 Arousal	-0.17	1.64	-0.11	0.918
T0 Dominance	0.21	1.12	0.19	0.854
T0 Valence	0.31	1.46	0.21	0.836
ΔT1 Arousal	0.38	1.62	0.23	0.821
ΔT1 Dominance	-0.61	1.63	-0.38	0.713
ΔT1 Valence	-0.10	0.92	-0.11	0.912
ΔT2 Arousal	-0.96	1.01	-0.94	0.360
ΔT2 Dominance	-0.81	0.66	-1.23	0.239
ΔT2 Valence	1.86	1.02	1.82	. 0.089
CTL1	-0.35	0.45	-0.78	0.445
CTL2	0.02	0.28	0.08	0.938
Teacher				
Intercept	4.51	0.99	4.58	*** <0.001
T0 Arousal	0.13	0.25	0.50	0.621
T0 Dominance	-0.18	0.48	-0.36	0.720
T0 Valence	-0.02	0.33	-0.07	0.945
ΔT1 Arousal	0.15	0.25	0.59	0.560
ΔT1 Dominance	0.36	0.42	0.85	0.402
ΔT1 Valence	-0.08	0.28	-0.29	0.774
ΔT2 Arousal	-0.52	0.48	-1.07	0.295
ΔT2 Dominance	0.22	0.52	0.42	0.676
ΔT2 Valence	0.002	0.25	0.01	0.992
CTL1	-0.42	0.22	-1.91	. 0.066
CTL2	0.43	0.21	2.08	* 0.047

Watertank				
Intercept	7.38	2.17	3.40	** 0.001
T0 Arousal	0.69	0.29	2.43	* 0.019
T0 Dominance	-1.78	0.74	-2.39	* 0.021
T0 Valence	0.28	0.20	1.37	0.178
ΔT1 Arousal	0.46	0.31	1.50	0.139
ΔT1 Dominance	-1.88	0.74	-2.54	* 0.014
ΔT1 Valence	0.93	0.29	3.17	** 0.003
ΔT2 Arousal	0.09	0.28	0.34	0.739
ΔT2 Dominance	-1.37	0.69	-1.99	. 0.052
ΔT2 Valence	0.52	0.21	2.48	* 0.017
CTL1	0.12	0.19	0.61	0.545
CTL2	-0.34	0.22	-1.59	0.117

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, . $p<0.1$

3.4 DISCUSSION

The first hypothesis states that work content influences the cognitive and affective factors. This hypothesis was divided over two aspects of work content; higher level goals consisting of different mission phases, and lower level goals, consisting of different tasks. The findings provided support for these hypotheses since the CTL and ES values varied significantly between different tasks and different phases of the Mars simulation.

It was expected that tasks of which the task goals had a good fit with the higher level goals, caused higher emotional state and cognitive task load levels. Following these expectations, the COLT tasks are expected to have higher CTL and ES levels since these tasks have goals fitting in the Mars mission environment. These expectations are found in the higher CTL levels for the teacher task compared to the Lunar Lander and Colored trails tasks. The Watertank task was not a realistic task and might therefore score lower on several valence measures compared to Lunar Lander and Cardiopres. The Cardiopres task on the other hand, had lower arousal and dominance than the Watertank task.

During different phases of the mission, cognitive and affective states had different values. At three instances the directions of these changes were as predicted. For example, the findings confirm that arousal decreased after the landing, suggesting adaptation or maybe boredom, between phases two and three (at T0 and T2). Next, after the initial adaptation and the prospect of landing on Mars, valence went up between phase one and two (at T0). Besides the findings for phase effects, the results revealed several other differences over non-adjacent phases. Important to note is that none of these findings contradicted hypothesised directions. For example, the perceived

level of cognitive task load (T1) was higher in the fourth phase compared to the second phase suggesting that tasks are seen as more difficult or more demanding during the last months of this simulated mission.

The second hypothesis stated that the cognitive and affective factors are predictors for task performance. The strongest support for this hypothesis was provided by the findings of the Watertank task, which showed valence as a significant predictor after Bonferonni corrections were applied to the results.

These findings support the thought explained by Matthews, Davies, Westerman, and Stammers (2008b), that stress reactions are distinguished on the characteristics of the challenge that is faced. Task characteristics determine the task goals and the fit with higher-level goals determine the appraisal of the task and emotional states. Matthews et al. (2008b) investigated the concept of Lazarus and Folkman and found that changes in stress state induced by tasks, varied with task demands. Matthews et al. (2008b) also state that subjective responses are influenced by a person's appraisal of the task and the environmental demands. When a task is appraised as overloading it will evoke stress, but when a task is appraised as a challenge it will evoke task engagement.

While the COPE model looks at a person's appraisal when faced with a stressful situation, one's motivation to solve a problem or execute a task is left out. However, literature explains that optimal performance on a task is also related to the amount of interest someone has for performing that task (O'Keefe & Linnenbrink-Garcia, 2014). Particularly the differences between the Cardiopres and Watertank exercises in the COLT-task can be explained by differences in experienced challenge and interests. Whereas the Cardiopres is a real (existing) payload for manned space missions, the Watertank was a simplified simulation of a complex system which the crew-members did not experience as realistic (e.g., leading to a different valence-arousal relationship).

Effects of work content on cognitive and affective processes and their effect on performances were found, even though the environment contained noise. The COPE model seems therefore a good starting point to dynamically adjust the work content to (predicted) cognitive, affective or performance factors. In the space domain, we are improving ePartner's (objective) real-time monitoring functions of crewmember's cognitive task load and emotional state, which is expected to provide better predictions for the task performance. In the space and naval domain, COPE-based support functions, like real-time feedback on the emotional state and error risks, are being designed and tested.

To appreciate the results of this study, a number of limitations should be considered. (1) Only six participants performed the tasks and they were selected on specific qualities as described in the method section. A previous study on the same dataset revealed that there might be cultural differences between the two groups of three participants (Smets et al., 2012). They did, however, participate in many sessions and thus provided a large set of data. To overcome any of these biases, all analyses accounted for individual user variation (Mirman et al., 2008). (2) The emotional state

and cognitive task load measures were all subjectively measured. The results can therefore contain social desirable answers and noise due to momentary perception biases. Since the COLT task asked for the teacher to rate the student and vice versa, the results of the COLT tasks might be biased. Neerincx, Kennedie, Grootjen, and Grootjen (2009) created predictive models with objective cognitive task load predictors and reached a higher accuracy in those predictions for the crewmembers of a naval Ship Control Centre (i.e., from 74% to 86% accuracy). Not only did they not experience social desirable answers, the data was also collected during the task and not afterwards. (3) The COPE model looks at the current state of a person when engaged in a stressful task. However, in this study participants were asked to rate their state in the task afterwards. (4) Although the findings did show significant regression models to explain task performance, for one of the collaborative training task, i.e. Cardiopres, and the negotiation and collaboration learning game, i.e. Colored Trails, the COPE factors did not lead to a significant regression model.

3.5 Conclusion

In the present chapter, data from the Mars500 project was fitted to the COPE-model to investigate influences on cognitive and affective measures, and those factors' predictability on task performance. The Mars500 project, a simulated Mars mission that lasted for 520 days, inflicts unique stressors such as social isolation, incidents and boredom on its crewmembers. Different tasks were performed during this mission, while gathering subjective emotional state and cognitive task load data.

The findings support the general hypothesis that work content, i.e. different tasks and mission phases, can influence cognitive and affective factors, and that these factors, on their turn, can explain task performance. Designing work for, or adjusting work plans during, long-term missions could benefit from this insight by considering cognitive task load and emotional state when (re)scheduling tasks. The findings also give some insight in the validity of the COPE model, showing the relation between external work factors, internal cognitive and cognitive factors and eventually the external performance when operating under stress.

3.6 Acknowledgements

This research (project number 056-22-010) is supported by the Dutch FES program "Brain and Cognition: Societal Innovation". Part of this research was conducted for the MECA project. This is a development funded by the European Space Agency (contract numbers 19149/05/NL/JA and 21947/08/NL/ST).

REFERENCES

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavioural therapy and experimental psychiatry, 25(1), 49-59.

Cohen, I., Brinkman, W. P., & Neerincx, M. A. (2012). Assembling a synthetic emotion mediator for quick decision making during acute stress. Paper presented at the Proceedings of the 2012 European Conference on Cognitive Ergonomics, Edinburgh.

Diggelen, J. v., & Neerincx, M. A. (2010). Electronic partners that diagnose, guide and mediate space crew's social, cognitive and affective processes. Paper presented at the Proceedings of Measuring Behaviour, Wageningen, The Netherlands.

Driskell, J. E., & Johnston, J. H. (1998). Stress exposure training. In J. A. Cannon-Browers & E. Salas (Eds.), Making decision under stress: Implications for individual and team training (Vol. 3, pp. 191-218). Washington, DC: American Psychological Association.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. The Annals of Statistics, 33(1), 1-53.

Gorbunov, R., Barakova, E. I., Ahn, R. M. C., & Rauterberg, G. W. M. (2011). Monitoring Interpersonal Relationships through Games with Social Dilemma. Paper presented at the IJCCI (ECTA-FCTA).

Gushin, V. I., Kholin, S. F., & Ivanovsky, Y. R. (1993). Soviet psychophysiological investigations of simulated isolation: some results and prospects. Advances in space biology and medicine, 3, 5-14.

Hennes, D., Tuyls, K., Neerincx, M. A., & Rauterberg, G. W. M. (2009). Micro-scale social network analysis for ultra-long space flights. . Paper presented at the IJCAI-09 Workshop on Artificial Intelligence in Space, Pasadena, California, US.

Keinan, G., Friedland, N., & Ben-Porath, Y. (1987). Decision making under stress: scanning of alternatives under physical threat. Acta psychologica, 64, 219-228.

Kieras, D., & Polson, P. G. (1985). An approach to the formal analysis of user complexity. International journal of man-machine studies, 22(4), 365-394.

Manzey, D. (2004). Human missions to Mars: new psychologica lchallenges and research issues. Acta Astronautica, 55, 781-790.

Matthews, G., Davies, D. R., Westerman, S. J., & Stammers, R. B. (2008). Stress, arousal and performance: an introduction Human Performance: cognition, stress and individual differences. New York: Psychology Press.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. Journal of memory and language, 59(4), 475-494.

Mundfrom, D. J., Jamis, J. P., Schaffer, J., Piccone, A., & Roozeboom, M. (2006). Bonferroni Adjustments in Tests for Regression CoefficientsMultiple Linear Regression Viewpoints (Vol. 32).

Neerincx, M. A. (2011). Situated Cognitive Engineering for Crew Support in Space. Personal and Ubiquitous Computing, 15(5), 445-456.

Neerincx, M. A., Bos, A., Olmedo-Soler, A., Brauer, U., Breebaart, L., Smets, N., . . . Wolff, M. (2008). The Mission Execution Crew Assistant: Improving Human-Machine Team Resilience for Long Duration Missions. . Paper presented at the Proceedings of the 59th International Astronautical Congress (IAC2008), Paris, France.

Neerincx, M., Kennedie, S., Grootjen, F., & Grootjen, M. (2009). Modelling the Cognitive Task Load and Performance of Naval Operators. Paper presented at the Lecture Notes in Artificial Intelligence. Schmorrow, DD; Estabrooke, IV; Grootjen, M.(Eds.), Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience. Proceedings of the 5th International Conference of the Augmented Cognition.

O'Keefe, P. A., & Linnenbrink-Garcia, L. (2014). The role of interest in optimizing performance and self-regulation. Journal of Experimental Social Psychology, 53, 70-78.

Ozel, F. (2001). Time pressure and stress as a factor during emergency egress. Safety science, 38, 95-107.

Smets, N. J. J. M., Cohen, I., Neerincx, M. A., Brinkman, W. P., & Diggelen, v. J. (2012). Improving crew support methods in human-machine teams for long-durations missions. Paper presented at the International Astronautical Congress, Naples, Italy.

Starcke, K., & Brand, M. (2012). Decision making under stress: a selective review. Neuroscience and Biobehavioral Reviews, 36, 1228-1248.

Sutcliffe, A., Ryan, M., Doubleday, A., & Springett, M. (2000). Model mismatch analysis: towards a deeper explanation of users' usability problems. Behaviour & information technology, 19(1), 43-55.

4. MODELLING ENVIRONMENTAL AND COGNITIVE FACTORS TO PREDICT PERFORMANCE IN A STRESSFUL TRAINING SCENARIO ON A NAVAL SHIP SIMULATOR

ABSTRACT

Professionals working in risky or emergency situations have to make very accurate decisions, while the quality of the decisions might be affected by the stress that these situations bring about. Integrating task- and bio-feedback into computer-based training environments could improve trainees' stress-coping behaviour. This chapter presents and assesses a refined version of the Cognitive Performance and Error (COPE) model that describes the effects of stressful events on decisions as a foundation for such a support tool. Within a high-fidelity simulator of a ship's bridge at the Royal Netherlands Naval College (RNNC), students of the naval college (n = 10) were observed while completing a 2 hour long shadowing and boarding operation combined with a search-and-rescue operation. For every action, variables were measured: objective and subjective task demand, challenge and threat appraisal, and arousal based on heart rate and heart rate variability. The data supported the COPE-model, and were used to create predictive models. The variables could provide minute-by-minute predictions of performance that can be divided into performance rated by experts and errors. The predictions for performance rated by experts correlated with the observed data (r = 0.77) and 68.3 % of the predicted errors were correct. The error predictions concern the chances of making specific errors of communication, planning, speed and task allocation. These models will be implemented into a real-time feedback system for trainees performing in stressful simulated training-tasks.

Keywords: stress, virtual training, cognitive errors, performance, simulator, navy

(Parts of) this chapter is published as: Cohen, I., Brinkman, W. P. & Neerincx, M. A. 2015. Modelling environmental and cognitive factors to predict performance in a stressful training scenario on a naval ship simulator. Cognition, Technology & Work. 1-17.

4.1 INTRODUCTION

Professionals working in safety-related fields such as the police force, fire department, aviation and the army, may enter uncertain and unexpected situations that bring along high levels of stress and demands (Driskell & Johnston, 1998). For example, naval ship operators encounter situations where they have to process a great amount of complex information in a short period of time and make a decision that can have severe consequences. Unfortunately, high levels of stress can negatively affect cognitive functions that are needed to execute several cognitive processes (Mendl, 1999). For example, errors are likely to occur in cognitive functions such as: attention, memory formation, and memory recall (Kleider, Parrott, & King, 2010; Mendl, 1999; Orasanu & Backer, 1996). In order to mitigate negative effects of stress, it is important to understand (1) the underlying processes and their effects on performance, and (2) the experiences with decision support systems that have been developed to improve performance. Understanding these two topics will help to achieve the aim of this chapter: Establishing predictive models that can be used in a new decision or training support system. This introduction starts with an overview of the literature on decision making under stress. Next, past and current decision support systems and other training methods are discussed to give an idea on what is important when designing such a system. The introduction then ends with a more detailed formulation of the research aim and hypotheses of this chapter.

Decision making involves a specific *cognitive process* that is influenced by high stress levels (Kerstholt, 1994; Starcke & Brand, 2012). Considering alternative decision options is a step in the decision-making process where stress can have negative effects. Individuals are more likely to decide without considering all alternatives (*premature closure*), use a non-systematic manner to consider the alternatives (*non-systematic scanning*), and seem unable to allocate time to consider all the alternatives (*temporal narrowing*) (Keinan et al., 1987). Time constrains seems to play a key role in these circumstances. For example, Maule, Hockey, and Bdzola (2000) found that time-pressure induced feelings of being energetic and anxious in people. But time pressure is not a prerequisite for stress. Keinan et al. (1987) reported that people can show disorganized and incomplete scanning when time limits are not present. Another observation relevant to these situations is that making a decision should not be seen as a single action, but as a chain of unfolding events and decisions. Ozel (2001) mentioned that human behaviour seems to be episodic in stressful and dangerous events. Every episode focuses on a certain goal that needs to be reached by executing appropriate actions. Achieving the goals can be seen as 'decision making between episodes' and achieving the actions can be seen as 'decision making within episodes'. Distinguishing goals and actions in human behaviours during emergency handling, makes it easier to investigate where in the decision processes stress plays a role. Another aspect of professionals working in stressful environments is that professionals often operate in teams. Working in a team can have obvious benefits, but also brings along extra cognitive issues that can have negative effects on performance during team decision-

making. Dowell and Hoc (1995) group these cognitive issues of coordinating decision-making and actions in four groups: planning, action, communication and task knowledge.

Current practices aiming to reduce negative effects of stress, make use of technical advances such as *decision support systems* or training environments that induce stress. Since the early 80's, research has tried to create effective digital decision support systems, or Intelligent Decision Aids (IDA) (Kontogiannis & Kossiavelou, 1999). Early support systems were designed to create decisions without biases. These systems provided limited options for the users to assess system's outcome: the users could merely accept or reject the decision made for them. This might have been a reason that the users had problems accepting these kind of decisions and support systems (Kontogiannis & Kossiavelou, 1999). Other problems were that the decision tools, even when focussed on naturalistic decisions, rarely showed decision improvement because individuals using them were often ahead of the tool (Cohen, 1993), and the tool designers cannot anticipate all possible scenarios that might occur (Reason, 1987). Therefore, recent and current IDAs are being designed to collaborate with its users to reach decisions, e.g., aiming at a "joint (human-technology) cognitive system" (cf.(Hollnagel & Woods, 2005)). In their review, Kontogiannis and Kossiavelou (1999) also propose that IDAs should try to prevent and delay stress. This can be done by implementing suggestions for changes in team strategies proven to be efficient while working under stress, into IDAs. IDAs should provide insight in event escalations and the anticipation of rare events. They should point out changes in communication necessary to work under stress and help the team members to keep track of each other's activities. Also the structure and task allocation of teams should adapt to stressful situations.

Another approach to prepare professionals to stressful enviroments, is to expose them to stressful conditions during *scenario-based training*, so that they can learn to cope with such conditions and to keep their performances at a high level in a stressful environment (Driskell & Johnston, 2006; Peeters et al., 2014). Previous research has found several aspects that can be applied to create effective stress-training. First, training environments should clearly convey a naturalistic environment. Making decisions in a real-life event is hardly the same as making decisions in a laboratory setting on which the classical decision theory is based (Beach & Lipshitz, 1993). Orasanu and Connolly (1993) listed eight factors that have been ignored in decision research, but are clear features of decision making in a naturalistic environment. The factors they list are: ill-structured problems, uncertain dynamic environments, shifting or competing goals, action or feedback loops, time stress, high stakes, multiple players and organizational goals, and norms. The presence of several of these factors in stressful situations will complicate the task of making a decision. Besides properties of naturalistic environments, specific guidelines have been suggested with regard to simulation training. For example, Sime (2007) listed seven properties for simulation training that help to reduce stress and its negative effects on decisions. Her seven

suggestions are: (1) when training certain skills that are to be applied in a stressful environment, the training setting should be a stressful environment as well; (2) reducing workload caused by time pressure can be achieved by rehearsing cognitive and behavioural skills up to automation and (3) by training heuristics of task prioritization; (4) cognitive rehearsal of a task can help increase one's confidence and ability (5) while team training increases team performance through the sense of team-identity; (6) changing the training environments helps train flexibility which makes it easier to work in an unknown situation; and last, (7) negative emotions and fear of the unknown can be reduced with the right training such as biofeedback training and cognitive control strategies.

Besides training in naturalistic environments, Sime (2007) suggested that *biofeedback* can be an effective tool to decrease stress during training. Whereas biofeedback increases control over one's physiological stress reactions (Bouchard et al., 2012), e.g., increased heart rate and fast respiration, cognitive control strategies can reduce emotions and distracting thoughts (Driskell & Johnston, 2006; Sime, 2007). Having a clear understanding of one's emotions will help individuals to experience fewer cognitive difficulties. It is argued that when under stress, cognitive attention resources will not only be depleted by the task at hand, but also by the emotional reactions (Driskell & Johnston, 2006; Gohm, Baumann, & Sniezek, 2001). When less cognitive resources are available, performance will decline. In other words, a better insight to one's emotional reactions improves performance under stress.

The project 'better decisions under high pressure' was started to develop computer-based training support for mitigating negative effects of stress on decision making. The envisioned support tool incorporates above-mentioned training and biofeedback approaches, i.e., by *combining* biofeedback (Sime, 2007), and suggestions for changes in strategies (Kontogiannis & Kossiavelou, 1999) and cognitive control strategies. Using only biofeedback teaches individuals to control their physiological reactions to stress, but not their cognitive reactions (Gohm et al., 2001; Keinan et al., 1987; Mendl, 1999). Cognitive feedback by suggesting efficient team strategies, together with biofeedback, could help trainees to overcome cognitive issues or impairments due to stress. In addition, it is expected that a tool that provides such combined support will be accepted better by the end-users.

To establish the real time bio- and performance-feedback, a model is needed that assesses the task and emotional load and provides performance predictions. The first model development step is to combine situational factors and cognitive and physiological indicators in a descriptive model and, subsequently, to refine it into a predictive model for cognitive processes and performances that are likely to occur in certain stressful situations. Cohen et al. (2012) provided a first (descriptive) version of this model based on literature on cognitive reactions to stress, called the COgnitive Performance and Error (COPE) model. The goal of this study is to validate a refined version of the COPE model, and test its ability to predict cognitive errors and performance. This chapter describes the acquisition of training data and the subsequent

analysis of the relationships between the COPE variables. The first hypothesis states that the variables are related as suggested by the COPE model. The second hypothesis states that the cognitive and situational variables in the COPE model can be used to predict performance and errors under stress. The next section of this chapter will describe the variables of the COPE model and their expected relations.

4.2 COPE-MODEL

The graphical representation of the COPE model displayed in figure 4.1 shows a cognitive process of decision-making under stress (Cohen et al., 2012). It roughly consists of three components: the work content, the individual's cognition and affect, and the individual's actions interpreted as the performance on a task or decision.

In this model, the work content consists of an event and the corresponding goals and objective task demand. An event itself is not stressful, but an individual can experience an event as stressful. Whether an event is experienced as a stressful one or not, depends on the individual's cognitive perception of the event. The task demand variables are based on Neerincx's (2003) model of Cognitive Task Load. In this model, task demand is divided into three dimensions: level of information processing, time occupied and task-set switches. By measuring these three dimensions, it is possible to determine cognitive task load during a specific task. The distinction between objective and subjective task demand implies that task demands can be determined "from the outside", e.g. by external experts or task analysts, (called "objective") and by the task performers themselves (called "subjective"). The subjective task demands can be lower or higher than the objective task demands (Bosse, Both, Lambalgen, & Treur, 2008).

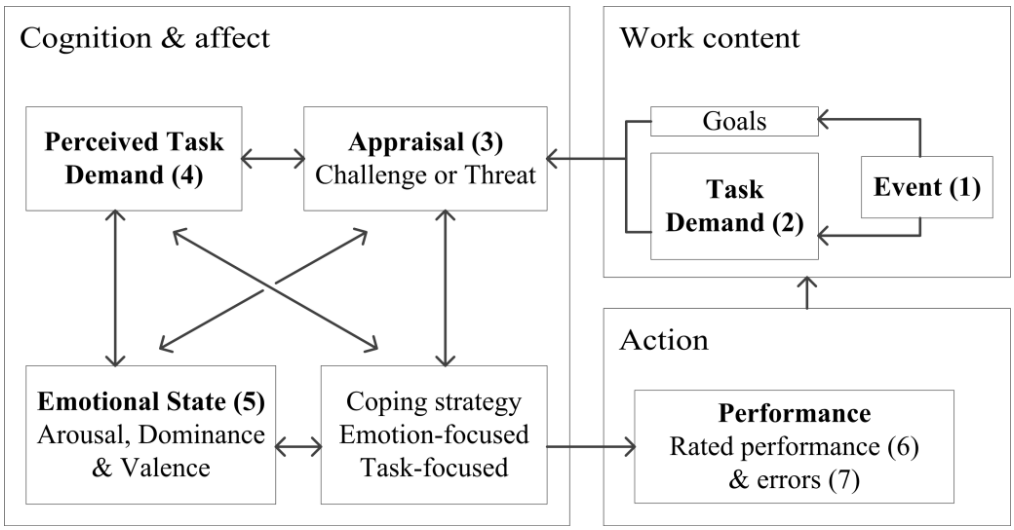


Figure 4.1. Schematic view of the COPE model of external and cognitive factors, predicting an individual's performance and errors.

Stress reactions that follow a stressful event can be explained as indirect reactions to the stressful event (Lazarus, 1999). After perceiving a stressful event, the severity of potential danger is assessed by the person experiencing it. This assessment is called the primary appraisal. If a situation is appraised as dangerous, it can be seen as a challenge when the individual feels he or she can cope with the event or as a threat when the individual feels he or she is lacking the resources to cope with the event. This is called the secondary appraisal. An individual that is experiencing a situation appraised as a threat or a challenge will try to cope with the situation by applying an appropriate coping strategy (Gaillard, 2007). Which coping strategy is used by the individual depends on the appraisal, but also on the individual's emotional state, since affect influences judgment (Forgas, 1995). The chosen coping strategy, in its turn, influences the decisions and actions made by the individual (Delahaij, 2009). Thunholm (2004, 2008) investigated individual's decision-making styles while under stress and found that an avoidant decision style relates to higher levels of distress and that a spontaneous decision style did not. Although decision-making styles and coping strategies fit in the COPE-model, they are out of the scope of this study, since there are no quick and easy ways to determine which style is used by the trainees.

A common way of measuring Emotional State is by using the valence, arousal and dominance scale (Bradley & Lang, 1994; Mehrabian, 1996). While valence is a scale that indicates the pleasantness of stimuli experienced by an individual, the arousal scale indicates the level of excitement. The dominance scale represents the level of control an individual feels. Instead of using a questionnaire, arousal can be measured in a less obtrusive way by measuring physiological aspects using biosensors (Haag, Goronzy, Schaich, & Williams, 2004). Physiological measures related to arousal induced by stress are, for example; heart rate (HR), heart rate variability (HRV) and stress hormone levels (Hjortskov et al., 2004; Krantz, Forsman, & Lundberg, 2004).

HR increases due to the Sympathetic Nervous System (SNS) stimulation caused, for example, by stress, exercise or cardiovascular disease. Activation of the Parasympathetic Nervous System (PNS) causes a decrease in HR. Changes in the balance between PNS and SNS activation produce heart rate fluctuations known as HRV. HR and HRV are used in the literature as measures of mental effort; an increase in mental effort will increase HR and decrease HRV (Mulder, 1992). Mulder (1992) describes a decrease in HRV as invested effort and not just a higher task difficulty. The effort needed to perform a more difficult task is shown by lowered HRV.

At the end of the cycle, an individual's cognition will lead to certain decisions and actions. Whether these decisions or actions are appropriate for the stressful event will determine the performance on the task. Reacting to the event will eventually result in changes of the external world and new tasks to perform and decision to make.

4.3 METHODS

After the explanation of the COPE-model in the previous sections, the hypotheses can be described in more detail. The first hypothesis states that the arrows in figure 4.1 represent correlations between the variables. The second hypotheses states that the cognitive variables (appraisal, task demand and physiological arousal) and the objective task demand can predict performance values.

To validate the COPE-model and use the variables to predict performance and cognitive errors under stress, seven variables from the COPE-model were measured (Section 3.3) while participants performed tasks in a stressful virtual scenario. The scenario took place in two simulated ship environments at the Royal Netherlands Naval College (RNNC) in Den Helder, the Netherlands. In every session, two teams of three participants were formed, each team in a separate simulator (simulators were connected). They experienced the same stressful scenario in which they needed to make decisions and execute tasks that would lead to a positive outcome.

4.3.1 PARTICIPANTS

Twenty-six students from the RNNC in Den Helder, The Netherlands, were recruited to participate in this experiment, including seven females. The median age was 22 years, with a minimum of 19 and a maximum of 41 years. Due to participant dropouts (caused by deployment, courses, etc.), two teams consisted of only 2 participants, and one session had only one team. Only participants with a complete data set, consisting of electrocardiogram (ECG) signals; questionnaires; and video data were included in the analyses. The final dataset consisted of 10 participants; two females and eight males of whom eight had between 0 and 2 years of operational service and two had over 2 years of operational service. The participants signed a consent form, and the study was approved by the ethical committee of Delft University of Technology, and the ethical committee of TNO.

4.3.2 MATERIALS

Two static bridge simulators from the RNNC were used: the primary simulator simulated the 'Hr. Ms. Tromp' frigate (figure 4.2), and the secondary simulator simulated the 'Hr. Ms. Van Amstel' frigate. These simulators consisted of a replica of the ships bridge and virtual surroundings, such as a moving horizon that gave the perception of ship movement. To control the ships, communication was necessary between the crew on the bridge (the participants), the superiors ashore (trainers), the crew on deck (trainers) and other ships (participants and trainers). In both simulators, at least two trainers were present during the scenarios.

4.3.3 MEASUREMENT OF VARIABLES

Seven variables were measured that appear in the COPE-model as appeared in figure

Figure 4.2. Bridge simulator based on the “Van Tromp” ship, seen from 2 angles and the trainer control room.

4.1: (1) events, (2) objective task demand, (3) appraisal, (4) subjective task demand, (5) emotional state/arousal, (6) performance, and (7) errors). For every event that occurred, these variables were measured. Since there were 21 identifiable events, it was not preferred to use long questionnaires since interruptions of complex tasks lowers their performance (Speier, Valacich, & Vessey, 1999). For coping strategy, no short questionnaire was found so a long questionnaire was used that measured general coping and not task specific coping. This questionnaire was filled in once. Therefore, the coping strategy measures were not used in the analyses. The different measurements are explained in the next subsections.

EXTERNAL WORLD: STRESSFUL EVENTS

A stressful, realistic scenario was written especially for this experiment by the simulator trainers of the RNNC. In table 4.1, the episodes, goals and actions of the tasks as suggested by Ozel (2001) are described. Five main episodes were identified: (1) shadowing the smuggling ship, (2) avoiding other vessels (this goal stays a goal during the whole experiment), (3) preparing for boarding, and (4) execute boarding and (5) reacting to and execute a search-and-rescue (SAR). Within these main episodes, different actions can be identified indicated in table 4.1 by the letters ‘a’ through ‘g’.

The scenario took place in the North Sea which is familiar territory for the participants. The scenario started with two navy warships shadowing a ship that was suspected of smuggling refugees. This ship discovered that it was being followed, which means they were likely to ‘destroy evidence’. In other words: throwing the refugees overboard. The participants needed to board the smuggling ship. Before the ship could be boarded, several actions needed to be taken. When the boarding was being executed, a Mayday call came in on the radio. The two Navy ships needed to decide to follow the distress call and transfer the boarding operation to another ship. When the Search-and-Rescue (SAR) was being executed, several actions needed to be taken. Depending on previous decisions and speed of the actions, some of the tasks could not be performed. All teams played the scenario for approximately 130 minutes.

Table 4.1. Actions that need to be executed in different stages of the scenario.

Episode	Time in scenario	Stressful events: actions for episode goal
1 Shadowing target ship	Start - ± 25 min	a. Start of the training b. Reacting when shadowing is discovered
2 Avoiding other vessels in the dark	During entire scenario	
3 Preparing to board target ship	± 25 min - ± 90 min	a. Deciding what team does what b. Positioning of the ships
4 Executing combined boarding	± 35 min - ± 90 min	a. Hailing of the target ship b. Positioning the target vessel c. Directing the crew d. Mutual communication e. Reacting on incoming Mayday
5 Executing Search and Rescue	± 90 min – end	a. Transfer target ship to arriving coastguard b. Launch helicopter c. Gearing up against traffic flow d. Navigate between sandbars e. Searching for 'man-over-board' f. Deploying the medic g. Carrying away injured

EXTERNAL WORLD: OBJECTIVE TASK DEMAND

Several questionnaires were available for measuring task demand. A reliable, fast and easy scale, is the Overall Workload questionnaire (Hill et al., 1992). This questionnaire consists of one scale, ranging from 0 to 100. A similar single-scale questionnaire was used in this study to measure task demand assessed by the trainers. They filled in the 10 point task demand scale for novice students (0-2 years of experience) and more experienced students (more than 2 years of service). Although the measurement itself is "subjective", the trainers rated the events as external and objective experts (i.e., not participating in the stressful situation) from the trainees point of view. It was therefore used as measure for objective task demand as described in Section 2.

COGNITION: APPRAISAL

For every event in the scenario (table 4.1) the participants filled in an appraisal questionnaire. One scale running from (1) challenge to (10) threat was filled in for every event in the scenario. With this scale, the appraisal could not be filled in as 0 but was always biased towards either challenge or threat. The scores were separated into two variables: challenge and threat. The challenge variable was created out of the scores

from 1 to 5 correspond to 'very challenging' (1) and 'little challenging' (5). A threat variable was created out of the scores from 6 to 10 where 6 corresponds to 'little threatening' and 10 corresponds to 'very threatening'. Appraisal scores 5, 4, 3, 2 and 1 were reversed to challenge scores 1, 2, 3, 3 and 5. The appraisal scores 6, 7, 8, 9, and 10 were converted to the threat scores of 1, 2, 3, 4 and 5. In this manner, the two appraisal variables could be compared.

COGNITION: SUBJECTIVE TASK DEMAND

The subjective task demand was measured with the same questionnaire as the objective task demand. A single scale, ranging from 1 'not at all demanding', to 10, 'very demanding' was filled in by the participants scoring their own (subjective) task demand.

COGNITION: EMOTIONAL STATE: AROUSAL

To measure the participant's arousal levels during the experiment without having the participants fill in a questionnaire, six mobi8 systems from TMSi (Enschede, The Netherlands) were used. These devices measure electrocardiographs (ECG), which can be translated into heart rate and heart rate variability. Each mobi8 has three sensors: one sensor was placed on the right collarbone, another sensor under the left ribs, and a ground sensor was placed on the right side, as shown in figure 4.3. To ensure that participants could walk around freely, they carried the mobi8 devices in a suitable case.

ACTION: PERFORMANCE

At the end of the experiment, the 'performance' was assessed by the trainers. All events from the session were rated on a 10 point scale for every participant. At least two trainers scored each participant, in order to create an averaged performance score. Z-scores were calculated for the performance rates, to extinguish possible trainer biases. Z-scores for a single participants performance rate were calculated with the mean and standard deviation of all the performance scores from all participants.

ACTION: ERROR

Two Sony HDR-CX300E cameras, a Sony handy cam DCR-SR55 and a Panasonic HDC-RM300 camera were used to record the activities in the simulators. Two cameras were placed in each simulator. The video data were used to define what situation and action occurred every minute. These videos were used to observe the trainer comments that could be used to determine if, when and what kind of errors occurred.

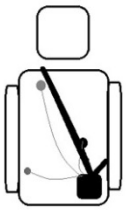


Figure 4.3. Sensor placement for ECG measurement with the Mobi8.

Within the video data, some errors were clearly identifiable. These errors were a direct result from faulty actions (Rasmussen, 1982). For example, in one team, the members were all focusing on their own task which made them forget to keep track of the radar and look outside. They did not notice a buoy in front of the ship, and navigated over the buoy. Only relative few of these kinds of errors happened during the experiment. Other errors were not directly visible, but the actions taken by the participants would not meet the planned goal. These actions would only unfold into an error, after a substantial amount of time had passed (Rasmussen, 1982). To identify these unfitting actions, the comments from the trainers were analysed. An example: the team members forgot to communicate their plan to the crew of the ship. If the crew does not prepare for action, the action cannot be performed when the participants want it to be performed. These errors, or more precisely, *tendencies to err*, were identified based on the comments and suggestions made by the trainers. Comments were categorized into 5 groups which corresponded to the groups of cognitive issues indicated by Dowell and Hoc (1995): communication; planning; speed; task allocation and 'other'. For every category, an example is given in table 4.2.

4.3.4 PROCEDURE

Five experimental sessions were performed. In an experimental session the scenario was played with six participants divided over two teams and simulators. The scenario lasted about 2 hours, with a 15 minute break halfway the scenario. Each team had a participant fulfilling the role of an *officer of the watch*, a *navigation officer* and a *steersman*.

Before running the scenario, the participants gathered in a classroom where they received a briefing about the scenario and the general aim of the study from the trainers. The participants were assigned to teams and divided the roles within the teams. After this, a questionnaire was filled in, in which general information about the participants was asked: e.g., years of service; experience in the simulators; and some general health questions, e.g., do you smoke, drink alcohol or caffeine. Next, the mobi8 systems were explained and connected to the participants. After the briefing, the participants went to the simulators were video cameras were turned on.

Table 4.2. Trainer comments can help in identifying the error category.

Category and description	Example
Communication: Participants forget to communicate information to other participants. This is a crucial point in co-operation.	The participants want to execute a boarding soon and they are informing the crew Trainer: *"You should not yet tell them about the boarding if it is not confirmed by the commander."*
Planning: When relevant information enters the bridge it can be used to make a plan for further actions. Often, participants have the information but have not made a plan yet.	The participants started a particular engine of the ship, which cannot run for longer than 15 minutes. Trainer: *"What are you going to do with these engines? They are going to break down soon."*
Speed: Speed is of major essence in this scenario. Plans need to be made fast, and actions need to be executed fast. The decision-making often takes too much time.	Between the ship and the mayday location are sandbanks. The students want to go around them. Trainer: *"Why do you want to go around them? Going between them is much faster."*
Task Allocation: Three people are on the bridge at all times (in this setting). They all have their own task, but when needed, task can be allocated differently to relieve one person of too much tasks. This is often forgotten.	One student is only focusing on reading the map.. Trainer: *"You should alternate between your tasks more."*

At the moment the simulators were started, the participants turned on the mobi8systems that started recording ECG. The first half of the scenario was played, followed by a 15-minute break, in which the participants answered the appraisal and task demand questionnaire for every action they encountered in the first half of the scenario.

The scenario was then continued. Due to differences in the decisions and actions taken by the different teams, not all sessions lasted the same amount of time. After approximately 2 hours, the scenario was ended by the trainers, and the second appraisal and task demand questionnaires were filled in about the events in the second half of the scenario.

The participants returned to the classroom where they took off the mobi8 sensors and were debriefed by the trainers. After the debriefing, the participants left and the trainers filled in the performance questionnaires, rating the actions of every participant. Although the basics of the scenarios were the same during every session, decisions made by the participants led to small differences in the storyline and the order of the events.

4.4 RESULTS

The results section is divided in two parts. The first part focuses on the variables of the COPE model. The second part focuses on creating a predictive model out of the data set. Before the data could be analysed, the raw measurements were first transformed into a data set ready for analysis.

4.4.1 DATA PREPARATION

The ECG data, as collected with the mobi8 from TMSi, were converted into heart rate (HR) and heart rate variability (HRV) per minute, using Matlab R2011a (The Mathworks). The signal measured in mV was first passed through a high-pass (0.5 Hz) and a low-pass (40 Hz) filter. After filtering, a peak-detection function was applied to the ECG signal. A minimum value had to be set in order to only detect the R-tops of the heart-beat. Counting the number of R-tops per minute resulted in the HR value per minute.

Nine outliers in the HR data, defined as values larger than three times the interquartile range, were removed from the data set, as they probably occurred because the heart-rate measurement devices had stopped, or were momentarily turned off. The HRV was calculated by the root mean squared successive differences (RMSSD) method. This method squares the average of the differences between two consecutive R-tops and was calculated for every minute.

A reliability analysis was conducted across the participants to examine similarity between participants' responses to their subjective task demand and appraisal. Table 4.3 shows the Cronbach's alpha values for both variables for the 26 participants and the group of n=10 from the final data set. Alpha values range from 0.75 to 0.99, it seems that there was a strong correlation between the participants' appraisal and subjective task demand.

Table 4.3. Cronbach's alpha for appraisal and subjective task demand scores between participants and subjective task demands scores between trainers.

	Cronbach's alpha	
	26 pp	10 pp
Appraisal	0.92	0.92
Subjective task demand	0.99	0.75

Table 4.4. A small part of the complete dataset. The columns indicate; participant, minute, heart rate, heart rate variability, appraisal (threat and challenge) task demand (objective and subjective) normalized performance and the error status (0 = No, 1 = Yes).

				Appraisal		Task demand			
pp	time	hr	hrv	threat	challenge	objective	subjective	Performance	Error
2	1	104.32	0.58	0	1	4.50	5	-0.79	0
2	2	97.75	0.61	2	1	8.50	13	-0.20	0
2	3	98.03	0.61	2	0	4.00	8	0.39	0
2	4	97.07	0.61	0	1	4.50	5	-0.79	0
2	5	99.73	0.60	0	6	5.67	0	0.00	0
2	6	101.65	0.59	0	6	5.67	0	0.00	0
2	7	97.72	0.61	2	6	9.67	8	0.19	0
2	8	104.82	0.57	2	0	4.00	8	0.39	1
2	9	101.16	0.59	0	6	5.67	0	0.00	0
2	10	107.49	0.56	0	7	10.17	5	-0.40	0

With the help of video data, it was determined which action (from table 4.1) was executed at which time by each participant. The comments from trainers were used to determine if errors were (almost) made by the participants. For every action, data about the appraisal, task demand, and performance were collected by means of the questionnaires described in the method section. Knowing what actions were executed every minute allowed us to calculate the appraisal, task demand, and performance per minute. If multiple tasks were performed in one particular minute, the associated appraisal and task demand scores were summed. For performance, scores were normalized and averaged per minute for all the tasks performed. Since the sessions all lasted over 2 hours, around 130 data points per participant were collected. As an example, a small part of the data set is displayed in table 4.4.

Table 4.5. Cohen's kappa for the inter-rater correlations between 3 raters and 5 categories

Category	coder1 coder 2	coder1 coder 3
Communication	1.00	0.77
Planning	1.00	0.72
Speed	1.00	0.80
Task allocation	1.00	0.76
Others	1.00	0.46

Besides the minute-by-minute data, six extra lag-variables were created for HR, HRV, threat, challenge, objective and subjective task demand, and the errors and performance variables. These lab-variables were created by taking the average value over a window of the previous five minutes. Using lag-variables might result in better predictions if the effects of stress are delayed or take more time to appear than one minute. For the error variable, the lag-variable would be a '1' if the previous 5 minutes would contain a '1'.

The trainer comments were coded by three independent coders, into five categories (table 4.2). Coder 1, the experiment leader, coded the comments into the five categories and made a description of the categories. These were explained to coder 2 and 3. The first round of codes were examined and the non-matching codes were discussed. Then, coders 2 and 3 coded the comments a second and a third time. As can be seen in table 4.5, coder 2 fully agreed with the coding of coder 1 while coder 3 had some disagreements. Table 4.5 shows the Cohen's kappa for inter-rater agreement. The inter-rater agreement ranges between 0.72 and 1, except for the 'other' category, that had the lowest inter-rater correlation of 0.46. This category was therefore left out of the analyses.

4.4.2 COPE MODEL EXPLORATION

The first step into the exploration of the COPE model was to examine the different variables and therewith testing the first hypothesis. Table 4.6 shows the sample size, minimum and maximum score, mean and standard deviation for each variable in the data set. There are less data points for the lag variables than for the non-lag variables, because the lag-variables were calculated starting at the sixth minute of the session.

After removing the heart rate outliers, the lowest heart rate recorded is 45.48 beats per minute, and the highest is 116.82 beats per minute. It is interesting to note that the mean of the normalized performance, lies below 0. The error scores are either one, or zero. The mean scores for all the error variables are close to zero, which illustrates an underrepresentation in the error data, which will be discussed later in this chapter.

Next, the correlations between the different variables were examined. To control for between participants variance, correlations of the variables were first calculated per participant and then averaged. The average amount of data points per participant is 116 which gives a df of 114. The critical correlation value for df = 114, and $\alpha = 0.05$ is rc = 0.179. Table 4.7 shows all the correlations. Highlighted correlations are significant correlations between different variables.

Table 4.6. Descriptive statistics of the model's variables and the lag-variables.

	N	Minimum	Maximum	Mean	Std. Deviation
Emotional State (arousal):					
Heart rate	1168	45.48	116.82	80.96	12.56
Heart rate variability	1168	0.51	1.38	0.76	0.14
Appraisal:					
Threat	1168	0	8	0.68	1.43
Challenge	1168	0	20	4.76	3.75
Task demand:					
Objective	1168	0	24.33	8.31	4.54
Subjective	1168	0	26.00	7.20	5.55
Actions:					
Performance	1168	-3.15	1.57	-0.45	1.06
Errors	1168	0	1	0.09	0.28
Communication	1168	0	1	0.04	0.19
Planning	1168	0	1	0.04	0.20
Speed	1168	0	1	0.01	0.12
Task allocation	1168	0	1	0.01	0.11
Other	1168	0	1	0.02	0.14
Lag variables:					
Heart rate	1119	50.50	109.59	81.01	12.06
Heart rate variability	1119	0.55	1.22	0.76	0.14
Appraisal threat	1119	0.00	6.20	0.69	1.29
Appraisal challenge	1119	0.00	15.98	4.71	3.32
Objective task demand	1119	1.80	18.93	8.29	3.48
Subjective task demand	1119	0.00	19.60	7.20	4.64
Performance	875	-3.15	1.57	-0.43	0.96
Error	1103	0	1	0.37	0.48

Table 4.7 shows a negative correlation between heart rate and heart rate variability; higher HR correlates to lower HRV and vice versa. This relations can be seen in both the 5 minute lag measure and the 1 minute measure. Among the regular variables, six significant correlations were found. Challenge and threat appraisals show a negative correlation as expected since appraisal was measured on a single scale

Table 4.7. Correlations between all the variables. Calculated by averaging the correlations between variables for every participant.

		Regular variables						Lag-variables						
		HR	HRV	Appraisal Threat	Appraisal Challenge	Objective Task Demand	Subjective Task Demand	HR	HRV	Appraisal Threat	Appraisal Challenge	Objective Task Demand	Subjective Task Demand	Performance
Regular variables	HR	1.00												
	HRV	-0.99	1.00											
	Appraisal Threat	0.03	-0.03	1.00										
	Appraisal Challenge	0.14	-0.13	-0.23	1.00									
	Objective Task Demand	0.09	-0.08	0.40	0.63	1.00								
	Subjective Task Demand	-0.01	0.02	0.47	0.23	0.75	1.00							
Lag variables	HR	0.44	-0.44	0.09	0.18	0.10	-0.03	1.00						
	HRV	-0.48	0.49	-0.09	-0.19	-0.09	0.06	-0.96	1.00					
	Appraisal Threat	0.03	-0.04	0.41	-0.17	0.07	0.14	0.07	-0.09	1.00				
	Appraisal Challenge	0.12	-0.12	-0.21	0.45	0.17	-0.10	0.23	-0.22	-0.27	1.00			
	Objective Task Demand	0.08	-0.07	0.03	0.21	0.37	0.22	0.14	-0.13	0.32	0.53	1.00		
	Subjective Task Demand	-0.02	0.04	0.13	-0.04	0.24	0.46	0.01	0.03	0.42	0.06	0.71	1.00	
	Performance	0.03	-0.03	0.02	0.08	0.03	0.05	0.03	-0.05	-0.10	-0.01	-0.03	-0.05	1.00
	Error	0.02	0.00	-0.03	-0.12	-0.10	-0.05	-0.01	0.01	0.08	-0.03	0.03	0.06	-0.05
Lag	Performance							-0.03	0.03	0.07	0.09	0.12	0.17	
Lag	Error							0.00	0.04	0.08	-0.24	-0.07	0.06	-0.00

Highlighted values are significant at α=0.05, n = 116, *df* = 114, r_c = 0.18

ranging from challenge to threat. Objective and subjective task demand correlated positively, indicating that participant and trainer perception corresponded to each other. Likewise, a positive correlation was found between task demands and both threats and challenge appraisals. This suggests that low task demand situations were not likely to be appraised as a threat or a challenge, while highly demanding situations were.

The correlations between the lag-variables show similar patterns, with two exceptions: a challenge appraisal was no longer found to correlate with subjective task demand, but was found to correlate positively with hearth rate and negatively with heart rate variability. In other words, this result supports the COPE model's link between arousal and challenge appraisal. Interestingly, no direct correlations were found between variables from the model and the minute-by-minute performance and errors (table 4.7). Still, on a five minute window, the lag variables show that challenge appraisal was reversely correlated with errors.

4.4.3 PREDICTIVE MODELS

Four Generalized Linear Mixed Model (GLMM) analyses were conducted to analyse the relation between the COPE model variables and the observed performance and cognitive errors. These analyses tested the second general hypothesis of this study. Performance and errors were modelled as dependent variables, using a linear model and a binary logistic regression model, respectively. The fixed factors consisted of the independent variables HR, HRV, threat, challenge, objective task demand and subjective task demand and their the lag-variables. 'Participant' was included as a random factor, thereby including a random intercept for each participant. The Variance Component type was used for random effect covariance type.

PERFORMANCE

A GLMM shows that the fixed factors can explain the performance per minute, ($F(6,1.161) = 8.60$, $p < 0.01$) with a correlation of $r = 0.77$ between observed and predicted performance. The individual variance differed significantly from the standard intercept ($var_{intercept} = 0.718$, *Std Err* = 0.35, $Z = 2.08$, $p. = 0.037$), indicating that on average the participants differed in their performance variance among each other. Examining the coefficients in table 4.8 shows that an increase in threat or challenge appraisal coincided with significant increase in the performance. The analysis show an opposite effect for objective task demand. An increase in this factor coincided with significant decrease in performance. Including the lag variables in the GLMM analysis resulted again in a model with explaining ability ($F(12,1.106) = 5.99$, $p<0.01$) with a correlation of $r = 0.79$ between predicted and objective performance. Also this model shows a significant random intercept for individual participants (varIntercept = 0.723, Std Err = 0.35 , Z = 2.06, p = 0.039). In addition to factors already found in the previous model, the extended model also revealed that an increase lagged threat appraisal of the last 5 minutes coincided with reduction in performance (table 4.9).

Table 4.8. Results of GLMM analysis on performance without lag-variables.

	df1	*df2*	*F*	*p*	Coeffi-cient	Std error	*t*	*p*	Lower	Upper
Corrected Model	6	1.161	8.60	** <0.01						
HR	1	1.161	0.18	0.68	0.00	0.01	-0.42	0.68	-0.02	0.02
HRV	1	1.161	0.78	0.38	-0.73	0.83	-0.88	0.38	-2.36	0.9
Appraisal Threat	1	1.161	20.46	** <0.01	0.12	0.03	4.52	** <0.01	0.07	0.18
Appraisal Challenge	1	1.161	33.67	** <0.01	0.07	0.01	5.8	** <0.01	0.05	0.09
Objective task demand	1	1.161	22.99	** <0.01	-0.06	0.01	-4.8	** <0.01	-0.08	-0.04
Subjective task demand	1	1.161	0.62	0.43	0.01	0.01	0.79	0.43	-0.01	0.03
Intercept					0.42	1.45	0.29	0.77	-2.42	3.26

. p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Table 4.9. Results of GLMM analysis on performance with lag-variables.

	df1	*df2*	*F*	*p*	Coeffi-cient	Std	*t*	*p*	Lower	Upper
Corrected Model	12	1.106	5.99	** <0.01						
HR	1	1.106	0.58	0.45	-0.01	0.01	-0.76	0.45	-0.03	0.01
HRV	1	1.106	0.72	0.407	-0.80	0.94	-0.85	0.40	-2.64	1.05
Appraisal Threat	1	1.106	34.64	** <0.01	0.19	0.03	5.89	** <0.01	0.13	0.26
Appraisal Challenge	1	1.106	21.14	** <0.01	0.07	0.01	4.60	** <0.01	0.04	0.10
Objective task demand	1	1.106	18.65	** <0.01	-0.06	0.01	-4.32	** <0.01	-0.09	-0.03
Subjective task demand	1	1.106	0.37	0.55	0.01	0.01	0.61	0.55	-0.02	0.03
Lag_hr	1	1.106	0.07	0.80	0.00	0.01	-0.26	0.80	-0.03	0.02
Lag_hrv	1	1.106	0.60	0.44	-0.86	1.11	-0.77	0.44	-3.03	1.32
Lag_Appraisal Threat	1	1.106	15.48	** <0.01	-0.17	0.04	-3.93	** <0.01	-0.25	-0.08
Lag_Appraisal Challenge	1	1.106	0.16	0.69	-0.01	0.02	-0.40	0.69	-0.04	0.03
Lag_Objective task demand	1	1.106	0.00	0.96	0.00	0.02	0.05	0.96	-0.04	0.04
Lag_Subjective task demand	1	1.106	0.18	0.67	0.01	0.01	0.42	0.67	-0.02	0.03
Intercept					1.78	1.83	0.97	0.33	-1.81	5.37

. p<0.1, * p<0.05, ** p<0.01, *** p<0.001

PREDICTIVE ERROR MODELS

The GLMM analysis revealed a significant binary logistic model for the error variable, $F(6,1.161) = 5.57$, $p. < 0.01$. On average, the model predicted 91.2% of the error status correctly, with 100% correct predictions for 'no error', and 0% correct predictions for 'error'. The model found no significant ($var_{intercept} = 0.195$, $Std.\ Err = 0.198$, $Z = .984$, $p = 0.33$) difference between the participants with regard to making an error. Table 4.10 shows that an increase in challenge appraisal coincided with an increased chance of making an error. Extending the model with lag variables resulted again in a significant model ($F(12,1.106) = 4.29$, $p < 0.01$) however without any significant coefficient (all $p > 0.05$).

Table 4.10. Multilevel linear regression for error prediction without lag-variables.

	df1	*df2*	*F*	*p*	Coeff.	Std	*t*	*p*	Exp Coeff.	Lower	Upper
Corrected Model	6	1,161	5.57	**<0.01							
HR	1	1,161	1.83	0.18	0.11	0.08	1.35	0.18	1.12	0.95	1.32
HRV	1	1,161	3.24	. 0.07	17.04	9.46	1.80	.0.07	<0.001	0.22	<0.001
Appraisal Threat	1	1,161	2.45	0.12	0.21	0.13	1.57	0.12	1.23	0.95	1.59
Appraisal Challenge	1	1,161	5.64	*0.02	0.15	0.06	2.37	*0.02	1.16	1.03	1.31
Objective task demand	1	1,161	3.08	.0.08	0.10	0.05	1.76	.0.08	1.1	0.99	1.22
Subjective task demand	1	1,161	1.60	0.21	-0.06	0.05	-1.27	0.21	0.94	0.86	1.03
Intercept					-20.51	13.79	-1.49	0.14	0.00	0.00	700.79

. p<0.1, * p<0.05, ** p<0.01, *** p<0.001

The analysis of errors led to two important observations: (1) as only 91.2% (1065/1168) of intervals included no error, the prediction was strongly biased towards no error prediction; and (2) no individual difference between participants were found. Such an error prediction model would not be useful in a training setting. Instead, in such a setting it would be acceptable to have some level of false alarms, if it would increase the number of correct predicted errors, i.e. hits. Therefore analyses were also conducted that corrected for the bias towards no error and no long include participants as a random intercept, i.e. a normal logistic regression was deemed sufficient.

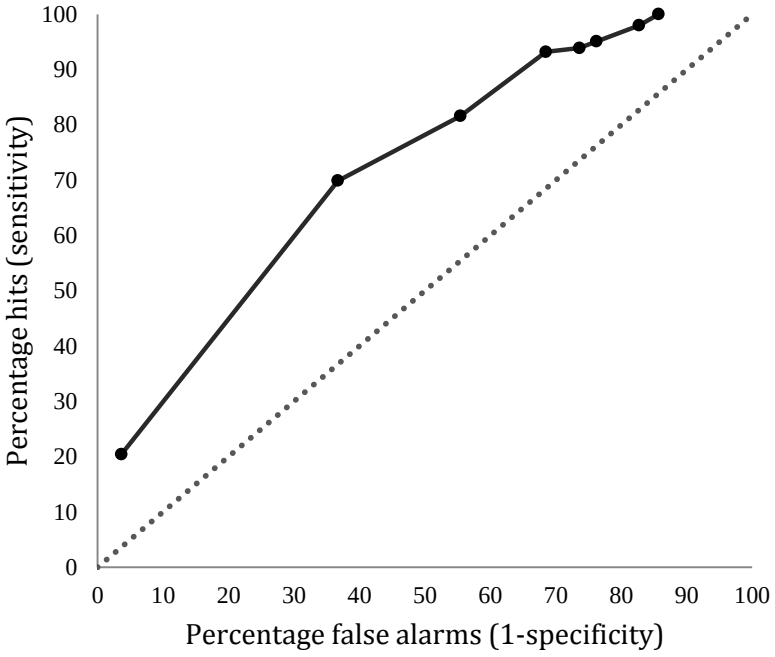


Figure 4.4. ROC curve consisting of logistic regressions for the error variable with different weighted cases

The underrepresentation of errors in the dataset was corrected by giving weights to the different cases. A receiver operation characteristic curve (ROC-curve) was used to determine the proportion for the case weighting that gives the most optimal logistic regression results. Figure 4.4 shows the ROC-curve with the false-alarm rate on the x-axis and the hit-rate on the y-axis. Two methods for determining the optimal weighting were used, namely the closest point to the ideal situation of 100% hits and 0% false alarm (d^2), and the maximum sum of sensitivity (Sn) plus specificity (Sp) (Kumar & Indrayan, 2011). Applying these methods, a weight ratio of 90:10 was determined for error versus no error, as this ratio resulted in a logistic regression with the highest sum of specificity plus sensitivity (1.33) and the shortest distance from the ideal left upper corner of ROC (distance = 0.23). Applying this weighing ratio led to significant logistic regression model ($\chi^2(6, n=1168) = 3761.26$, $p < 0.01$) that included an intercept and the other COPE model variables. Whereas the logistic regression model with only an intercept had a correct prediction rate of 53.5%, adding the variables improved this to 66.4%, with an Cox & Snell's $R^2 = 0.17$. Adding the lag-variables also created a significant model ($\chi^2(12, n=1168) = 4567.445$, $p<0.01$) with a correct prediction rate of 68.3% and a Cox & Snell's $R^2 = 0.21$. This model had a correct prediction of 52.3% when only the intercept was used. As Table 4.11 shows all the coefficient in the model are significant (all $p <.05$).

Table 4.11. Results of weighted logistic regression for the error variable including lag variables.

	B	*S.E.*	Wald	df	Sig.	Exp(β)
HR	-0.04	0.02	5.32	1	0.02	0.964
HRV	-7.61	1.76	18.66	1	<.01	4.95×10^{-4}
Appraisal Threat	-0.36	0.03	208.30	1	<.01	0.696
Appraisal Challenge	-0.14	0.01	139.39	1	<.01	0.868
Objective task demand	-0.07	0.01	36.99	1	<.01	0.937
Subjective task demand	0.03	0.01	13.81	1	<.01	1.034
Lag_hr	-0.21	0.02	119.41	1	<.01	0.813
Lag_hrv	-21.29	2.07	106.22	1	<.01	5.67×10^{-10}
Lag_Appraisal Threat	0.33	0.03	144.09	1	<.01	1.39
Lag_Appraisal Challenge	0.03	0.01	6.93	1	0.01	1.035
Lag_Objective task demand	0.08	0.01	43.09	1	<.01	1.079
Lag_Subjective task demand	0.02	0.01	4.72	1	0.03	1.021
Intercept	41.12	2.43	287.36	1	<.01	7.18×10^{17}

. p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Table 4.12. Logistic regressions for the four error categories with lag-variables.

	Optimal weight	d^2	Sn + Sp	Model	Correct Predictions [1]	Cox & Snell's R^2
Communication	98:02	0.141	1.48	χ^2 (12, n=1119) = 2288.835, p<0.05	51.3% – 74.1%	0.34
Planning	98:02	0.266	1.308	χ^2 (12, n=1119) = 1180.384, p<0.05	53.9% – 66.4%	0.19
Speed	99:01	0.194	1.385	χ^2 (12, n=1119) = 860.390, p<0.05	56.3% – 69.9%	0.24
Task allocation	99:01	0.017	1.822	χ^2 (12, n=1119) = 2219.745, p<0.05	64.3% – 91.5%	0.55

[1]Correct predictions for intercept model and for intercept + variables model

This same procedure was also used to conduct logistic regression analysis on the specific type of errors, i.e. communication, planning, speed, and task allocation. Table 4.12 shows the different weighting ratios used for each error category. The correct prediction ranged from 66.4% for planning errors to 91.5% for task allocation errors. All logistic regression models were significant ($p < .05$) with Cox & Snell's R^2 ranging from 0.19 to 0.55.

4.4.4 CROSS-VALIDATIONS

To test the generalizability of the performance model, a cross-validation was conducted (Refaeilzadeh, Tang, & Liu, 2009). This means that the dataset was divided in two sets: one to train the model and one to validate the model. The leave-one-out cross validation, a specific form of *k*-fold cross validation, was applied. Here, the data set was divided in 10 parts. Data from 9 participants was used as the training part to create the regression model, i.e. determine the coefficients. This would lead to formulas with a general form:

$$Predicted\ performance = intercept + (b * HeartRate) + (b * HeartRateVariability) + (b * Threat) + (b * Challenge) + (b * Objective\ Task\ Demand) + (b * Subjective\ Task\ Demand)$$

Data from the participant that was left out was used as the validation part of the model by entering the actual values of the predictors, included the lag-variables, and calculating the predicted performance. Every participant was used once as the validation part, which created predictive performance values for all the participants.

The predicted performance values from a GLMM (including lag variables) without random factors, also known as a linear regression, correlated with observed performance values (r = 0.56). A cross-validation for this model still showed a significant correlation, although reduced ($r(1168) = 0.17$, $p < 0.01$).

A similar procedure conducted for the weighted logistic regression model on the cognitive error in general, where the total logistic regression model (including lag variables) correlated with the observed errors with an r = 0.23, the cross-validation model lowered this correlation with the observed errors to $r(1165) = 0.13$, $p < 0.01$. This cross-validation model for the errors had a correct prediction of 67.3%, which is close to the 68.3% correct prediction for the model based on total sample.

4.5 Discussion and Conclusion

The first hypothesis of this study states that there are relationships between the variables of the COPE model. As the correlation table shows, correlations exist between the variables. Only the physiological variables of heart rate and heart rate variability do not seem to correlate to the other cognitive or performance variables.

The second hypothesis was also confirmed. Models were created that use situational and cognitive variables to predict performance and errors. Tables 4.8, 4.9 and 4.11 show the contribution of the variables to the different outcome variables. Figure 4.5 shows how performance and errors can be predicted out of the COPE-variables. The analyses used in this study showed how much of the variance in the performance and error variables was accounted for by the COPE model's variables. The significant predictions found in the analyses are presented as arrows in figure 4.5. Performance rated by experts can be predicted out of the threat, challenge and objective task demand variables (solid lines), but not out of the physiological measures of arousal. Participants walked around in the simulator and this might have been a distorting factor in the measurements of arousal. The strong correlation that was found between HR and HRV but not between HR or HRV and the other variables might show a ceiling effect.

Errors, on the other hand, could be predicted out of the physiological measures, which might indicate that the method used for scoring 'expert-rated-performance' might have been un-synchronized with the ECG measures. The errors were extracted every minute from the videos and are therefore better synchronized with the ECG measures that were also measured per minute. The performance scores were measured with questionnaires that listed the executed tasks in the same way as the questionnaires for appraisal and task demand did. Future studies should look into the combination of different measurement systems and how to improve synchronization between these different measurements.

Errors can be predicted out of all variables (dotted lines). The ability to predict errors varies between error categories, with planning errors having the lowest and task allocation errors having the highest correct prediction rates. Furthermore, the cross-

validation analysis showed the possibility of making a significant prediction for a new data set, suggesting the generalisation of these prediction models. Other studies have done similar research, but within different context and with different methods. As a first example, Neerincx, Kennedie, Grootjen, and Grootjen (2009) created a Naïve Bayesian Network to predict performance of naval operators. The COPE model includes Neerincx' model, addressing more factors and distinguishing several error types, and can therefore be used for training purposes. A second example is the Structural Equation Model of Kylesten (2013) that describes dynamic decisions-making on operative levels. Kylesten (2013) also used a descriptive model to describe dynamic decision-making and fitted data to this model. In contrast to the COPE-model, this model did not include an objective measure from an instructor, and no physiological measures were used.

This study has a number of limitations that should be noted. Although observation data were collected from 26 participants, only data of 10 participants were included in the analysis, giving this study a small sample size regarding the number of participants involved. When it comes to the amount of 1-minute observations, this study had a relatively large sample (n = 1168). This sample ratio seems appropriate as the focus of the work was not to examine performance and cognitive errors between individuals, but between different stressful situations within subjects. Cross-validation analyses showed a reduction in prediction accuracy compared to the GLMM models, but the predictions still correlated significantly with the observation data. This supports the prediction models' ability to generalize outside the sample of individuals included in this study. Another limitation was that the data were collected within teams, and therefore individual observations might not be completely independent. Future studies that include more individuals might consider to include different teams as a random factor in the analysis. Future studies might also consider the effect of different individual characteristics, as this study found that performance prediction differed between the participants. For the arousal measurements, other physical indicators such as galvanic skin response, might be more suitable for a setting in which physical movement is inevitable.

There are several ways to increase the prediction accuracy of the models. First, broadening the 1-minute interval prediction window, for example to 5 minutes, might lead to higher accuracy in the predictions. Compared to the 1-minute performance and error variables, correlations between the 5-minute performance and error variables and the other variables are slightly stronger, with one significant correlation. It might be easier to predict over a longer period of time, but for a fast-paced stressful training scenario it might not be appropriate to deliver feedback for a 5 minute period; hence, this chapter mainly had a minute-by-minute focus.

Second, in this study, cognitive errors were defined as an intervention or comment by the trainers. When exactly a trainer decides to intervene or make a comment, might vary and the predictions per minute are likely to be error prone. Therefore, when giving minute-by-minute error feedback in a training situation, giving

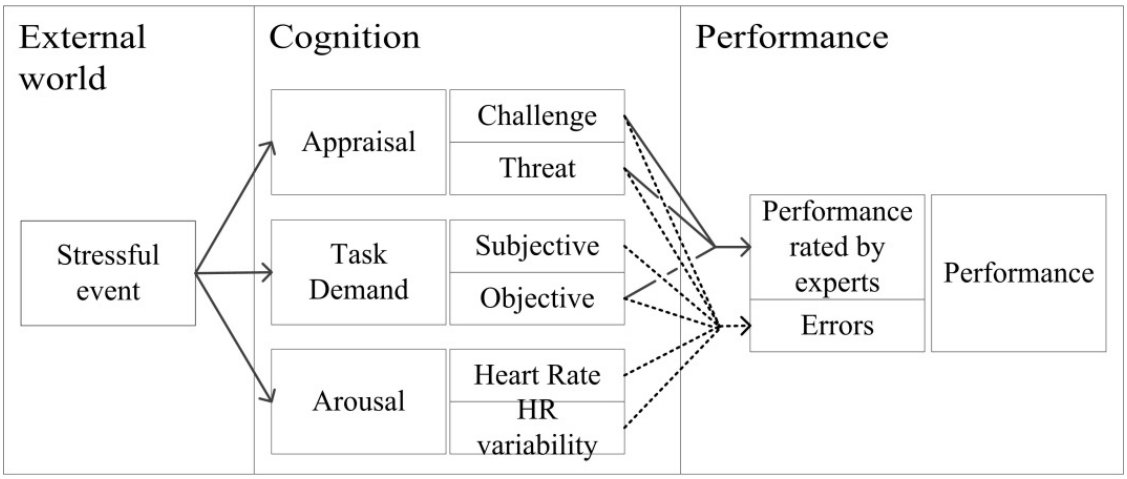


Figure 4.5. COPE model with indications of validated correlations, and validated predictive values.

error likelihood feedback might be more appropriate than a simple yes or no error type of feedback.

A third way to improve the models prediction accuracy might be to add information about the participants coping strategies. As can be seen in figure 4.1 and figure 4.5 coping strategy is an intervening variable between the other cognitive variables and the actions of the individual. According to the COPE model, the data used to predict the errors and performance were all indirect factors, and therefore less able to provide information for accurate prediction.

Besides the support found for the COPE-model, the second contribution of this chapter is the demonstration of creating a model for minute-by-minute predictions of performance and cognitive errors in a virtual stressful situation. When using such a model, the necessary information needs to be available per minute, in this case: the stressful environment, task demand, appraisal, and arousal. Arousal data could be obtained from physiological indicators. Assuming application of the models for the same training scenario as presented in this study, the same trainer data about the objective task demand could be used again. In an integrated environment, e.g. a virtual environment, a computer generates specific events in the training scenario, which provides the information about the stressful situation. Every event can be linked, for example, to a look-up table that holds the corresponding information about objective task demands for every event. In this study, the subjective task demand and appraisal information was obtained from students after completion of the scenario. For a minute-by-minute feedback system this would not be suitable, since the information is needed every minute. Asking the trainees to provide this information each time they are confronted with a new task would provide individual real-time information, but is too obtrusive and will lower the performance of the task (Speier et al., 1999), and affect their engagement or feeling of being present in such a situation (Hartanto et al., 2012) A less interruptive way would be to use the data provided by participants in this study as a

more general appraisal and subjective task demand. This last approach seems possible since high similarities were found between the participants' item responses (table 4.3).

The methods suggested in this chapter are in principle not limited to the training scenario used in this study. When applying it for other training scenarios, the variables related to the tasks (appraisal, task demand) need to be re-measured for every action or event occurring in that scenario. This will lead to new task coefficients that can be implemented in the created predictive models.

To conclude, the observational study and analysis presented in this chapter give an overview of which variables are important when making decisions in stressful situations and present a method to predict performance and errors from these variables. With the creation of predictive models, the next step is to implement them in a feedback system for training purposes as described in the introduction. Professionals would get real-time feedback on their expected performance and the possibility of making errors, based on their current state and the state of the external world. Training decision making under stress while receiving feedback would hopefully lead to an increase in performance and a diminishing of errors in real-live scenarios.

ACKNOWLEDGEMENT

The work presented in this chapter is supported by the Dutch FES program: Brain and Cognition: Societal Innovation (project no. 056-22-010). We would like to thank the the Royal Netherlands Naval College (RNNC) in Den Helder, The Netherlands. Especially the trainers at the bridge simulator for their help with running the experiment and gathering and scheduling the participants.

REFERENCES

Beach, L. R., & Lipshitz, R. (1993). Why classical decision theory is an inappropriate standard for evaluating and aiding most human decision making. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 3-20). Norwood, New Jersey: Ablex publishing corporation.

Bosse, T., Both, F., Lambalgen, R. v., & Treur, J. (2008). *An agent model for a human's functional state and performance.* Paper presented at the International Agent Technology, Sydney.

Bouchard, S., Bernier, F., Boivin, E., Morin, B., & Robillard, G. (2012). Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. *Plos one, 7*(4).

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavioural therapy and experimental psychiatry, 25*(1), 49-59.

Cohen, I., Brinkman, W.-P., & Neerincx, M. A. (2012). *Assembling a synthetic emotion mediator for quick decision making during acute stress.* Paper presented at the Proceedings of the 2012 European Conference on Cognitive Ergonomics, Edinburgh.

Cohen, M. S. (1993). The bottom line: naturalistic decision aiding. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 3-20). Norwood, New Jersey: Ablex publishing corporation.

Delahaij, R. (2009). *Coping under acute stress: the role of person characteristics.* Kon. Broese & Peereboom, Breda.

Dowell, J., & Hoc, J.-M. (1995). Coordination in emergency operations and the tabletop training exercise. *Le Travail Humain, 58*(1), 85-102.

Driskell, J. E., & Johnston, J. H. (1998). Stress exposure training. In J. A. Cannon-Browers & E. Salas (Eds.), *Making decision under stress: Implications for individual and team training* (Vol. 3, pp. 191-218). Washington, DC: American Psychological Association.

Driskell, J. E., & Johnston, J. H. (2006). Stress exposure training. In J. A. Cannon-Browers & E. Salas (Eds.), *Making decisions under stress* (Vol. 3). Washington, DC: American Psychological Association.

Forgas, J. P. (1995). Mood and judgement: the affect infusion model (AIM). *Psychological bulletin, 117*(1), 39-66.

Gaillard, A. (2007). *Stress productiviteit en gezondheid* (Vol. 3). Amsterdam: Holland Graphics.

Gohm, C. L., Baumann, M. R., & Sniezek, J. A. (2001). Peronality in extreme situations: thinking (or not) under acute stress. *Journal of research in personality, 35*, 388-399.

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion recognition using bio-sensors: first steps towards an automatic system *Affective Dialogue Systems, Tutorial and Research Workshop*. Kloster Irsee, Germany.

Hartanto, D., Kang, N., Brinkman, W.-P., Kampmann, I. L., Morina, N., Emmelkamp, P. M. G., & Neerincx, M. A. (2012). Automatic mechanisms for measuring subjective unit of discomfort. *annual review of cybertherapy and telemedicine*(181), 192-196.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human factors, 34*(4), 429-439.

Hjortskov, N., Rissen, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal Applied Physiology, 92*, 84-89.

Hollnagel, E., & Woods, D. D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering*. Boca Raton (Fl), USA: CRC Press, Taylor & Francis group.

Keinan, G., Friedland, N., & Ben-Porath, Y. (1987). Decision making under stress: scanning of alternatives under physical threat. *Acta psychologica, 64*, 219-228.

Kerstholt, J. H. (1994). The effect of time pressure on decision-making behaviour in a dynamic task environment. *Acta psychologica 86*, 89-104.

Kleider, H. M., Parrott, D. J., & King, T. Z. (2010). Shooting behaviour: how working memory and negative emotionaliy influence police officer shoot decisions. *Applied cognitive psychology, 24*, 707-717.

Kontogiannis, T., & Kossiavelou, Z. (1999). Stress and team performance: principles and challenges for intelligent decision aids. *Safety science, 33*, 103-128.

Krantz, G., Forsman, M., & Lundberg, U. (2004). Consistency in Physiological Stress Responses and Electromyographic Activity during Induced Stress Exposure in Women and Men. *Integrative Physiological & Behavioral Science, Vol*(2), 105-118.

Kumar, R., & Indrayan, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics, 48*, 277-287.

Kylesten, B. (2013). Dynamic decision-making on an operative level: a model including preconditions and working method. *Cognition, technology and work*(15), 197-205.

Lazarus, R. S. (1999). *Stress and emotion: a new synthesis*. New York: Springer Publishing Company, Inc.

Maule, A. J., Hockey, G. R. J., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: changes in affective state and information procession strategy. *Acta psychologica, 104*, 283-301.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology, 14*(4), 261-292.

Mendl, M. (1999). Performing under pressure: stress and cognitive function. *Applied animal behaviour science, 65*, 221-244.

Mulder, L. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology, 34*(2), 205-236.

Neerincx, M., Kennedie, S., Grootjen, F., & Grootjen, M. (2009). *Modelling the Cognitive Task Load and Performance of Naval Operators.* Paper presented at the Lecture Notes in Artificial Intelligence. Schmorrow, DD; Estabrooke, IV; Grootjen, M.(Eds.), Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience. Proceedings of the 5th International Conference of the Augmented Cognition.

Neerincx, M. A. (2003). Cognitive task load design: model, methods and examples. In E. Hollnagel (Ed.), *Handbook of Cognitive Task Design* (pp. 283-305). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Orasanu, J., & Connolly, T. (1993). The reinvention of decision making In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 3-20). Norwood, New Jersey: Ablex publishing corporation.

Orasanu, J. M., & Backer, P. (1996). Stress and military performance. In J. E. Driskell & E. Salas (Eds.), *Stress and human performance* (pp. 89-125). Mahwas, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Ozel, F. (2001). Time pressure and stress as a factor during emergency egress. *Safety science, 38*, 95-107.

Peeters, M., Van Den Bosch, K., Meyer, J.-J. C., & Neerincx, M. A. (2014). The design and effect of automated directions during scenario-based training. *Computers & Education, 70*, 173-183.

Rasmussen, J. (1982). Human Errors: A taxonomy for describing human malfunction in industrial installations. *Journal of Occupational Accidents, 4*, 311-333.

Reason, J. (1987). Cognitive aids in process environments: prostheses or tools? *International journal of man-machine studies, 27*, 463-470.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross Validation. In L. Liu & M. T. Ozsu (Eds.), *Encyclopedia of Database Systems* (pp. 6): Springer.

Sime, J.-A. (2007). *Designing emergency response training: seven ways to reduce stress.* Paper presented at the International Conference on Cognition and Exploratory Learning in Digital Age, Algarve, PT.

Speier, C., Valacich, J. S., & Vessey, I. (1999). The influence of task interruption on individual decision making: an information overload perspective. *Decision sciences, 30*(2), 337-360.

Starcke, K., & Brand, M. (2012). Decision making under stress: a selective review. *Neuroscience and Biobehavioral Reviews, 36*, 1228-1248.

Thunholm, P. (2004). Decision-making style: habit, style or both? *Personality and individual differences, 36*, 931-944.

Thunholm, P. (2008). Decision-making styles and physiological correlates of negative stress: is there a relation? *Cognition and Neurosciences, 49*, 213-219.

PART 2
IMPROVING PERFORMANCE UNDER STRESS

5. A COPE-BASED FEEDBACK SYSTEM

5.1 INTRODUCTION

The previous chapters of this dissertation focused on the establishment and validation of the COPE model. Based on the COPE model, significant models were created to predict expert rated performance levels and four categories of errors and error tendencies. This chapter describes how these predictive COPE models are combined with a biofeedback method and how they are implemented into one feedback system. The feedback system is thereby able to provide biofeedback, predicted performance feedback, and predicted error-chance feedback. The primary use of this COPE-based feedback system, or COPE-FB system, is to provide support to professionals training to work under stress.

Biofeedback methods can be applied to control and reduce physiological reactions to stress (Sime, 2007) and help to improve task performances (Bouchard et al., 2012; Prinsloo, Derman, Lambert, & Rauch, 2013; Prinsloo et al., 2011). Besides physiological reactions, stress also influences cognitive processes (Gohm et al., 2001; Keinan et al., 1987; Mendl, 1999). Biofeedback techniques could therefore be extended with feedback techniques that provide insight into the effects of stress on cognitive processes. The COPE model includes such processes, and, therefore, seems to be a good candidate to generate joint biofeedback and cognitive feedback. Gonzalez (2005); Lerch and Harter (2001) found that a combination of outcome performance feedback (overall indication of performance) and feed-forward ('what-if' feedback) increase task performance. Instead of outcome feedback, the COPE-FB system provides predicted

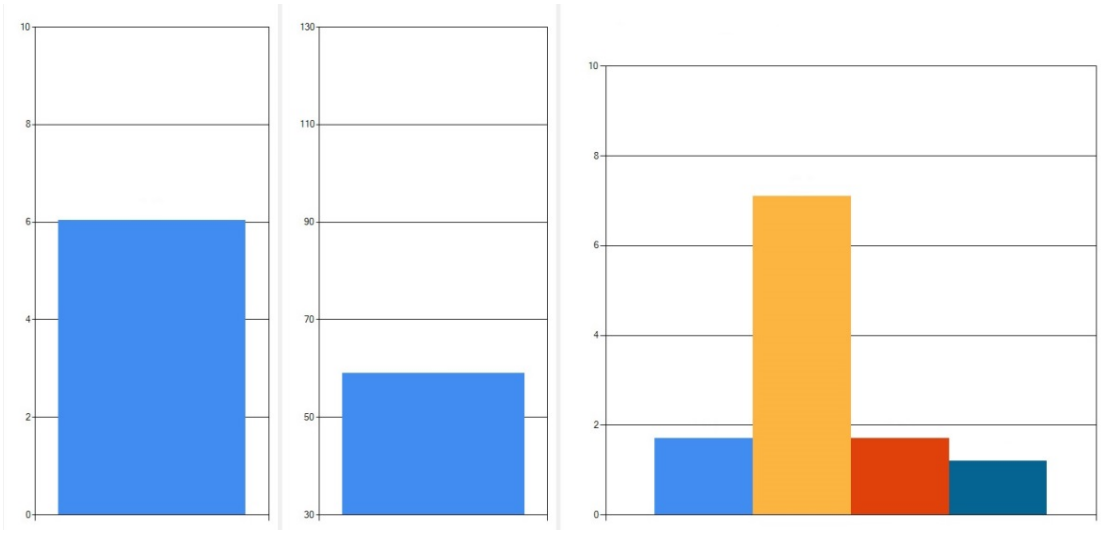


Figure 5.1 Simplified presentation of the feedback design of the COPE-FB system for, respectively, performance, physiological, and error chance feedback. More details are discussed further in this chapter.

overall performance, and instead of feed-forward, the COPE-FB system provides predicted error-chances. The underlying assumption is that making trainees aware of their error tendencies will help them to avoid making these errors (Dörner & Schaub, 1994).

Feedback from the COPE-FB system (see figure 5.1) is provided in a real-time fashion, because immediate feedback makes learning more efficient than delayed feedback in digital tutors (Anderson, Corbett, Koedinger, & Pelletier, 1995). When trainees are in the same state when the task is performed and when feedback is provided, they learn in a state-dependent manner (Kenealy, 1997). Delaying feedback until after the training or task has been performed might cause the benefit from the state-dependent learning to be lost. A downside of immediate feedback is that it might be distracting and will thus be ignored, rendering it ineffective (Shute, 2008; Wickens, Lee, Liu, & Becker, 2004). Therefore, the COPE-FB system presents the immediate feedback during a pre-determined time. When trainees are performing a task, they have a short period of time to see the feedback before new feedback is calculated.

COPE-FB provides three different feedback types (physiological feedback, predicted performance feedback, and predicted error chance feedback) with similar (consistent) bar graph presentations (figure 5.1). For all three feedback types, a higher bar graph represents a higher value. The bar graphs belonging to the same feedback type are grouped according to the principle of common region (Rock & Palmer, 1990).

5.2 FEEDBACK SYSTEM REQUIREMENTS

The COPE-FB system uses linear regression models to predict performance (as rated by experts) and logistic regression models to predict chances of specific errors. These models all need six parameter values and six input values analogous to the variables from the COPE model: appraisal (challenge and threat level), task demand and perceived task demand, and arousal (measured with heart rate and heart rate variability).

For every training scenario in which the COPE-FB system is used, a calibration session is needed to establish the model parameters and input values. During such a calibration session, appraisal and task demand are measured for every task present in the training. This results in an averaged value for those variables for every task. The arousal measures are not assessed beforehand since they are determined in real-time with physiological measurements using the HxM Zephyr heart rate belt. These values and the coefficients are listed in task and regression files, which are read by the COPE-FB system.

5.2.1 TASK FILES

A file containing the variable values for every task needed to be created. Figure 5.2a shows a sample of the task file that was created with data from Chapter 4 and is used in

the experiment of Chapter 7. It shows the first three tasks and the values for threat, challenge, task demand, and perceived task demand. As the file shows, the first task (starting the training) was generally appraised as less of a threat than the second task (avoiding vessels in the dark).

5.2.2 REGRESSION FILES

The data necessary to conduct regression models (the model parameters) were also stored into files. Figure 5.2b shows a sample of a regression file containing the regression model for performance as it was established in Chapter 4 and will be used in Chapter 7. As figure 5.2b shows, the performance regression contains a constant variable (-0.467) and constant parameters (scalars) for the heart rate (HR), heart rate variability (HRV), threat, challenge, task demand, and perceived task demand. Baseline values were subtracted from the heart rate and heart rate variability.

The COPE-FB system needs both files to calculate the predicted performances. For example, to predict performance for task 1, a regression formula is established using the scalars from figure 5.2b.

```
[{"task 1" : "Start of the training",
        "threat" : 1,
        "challenge" : 2.39,
        "task demand" : 4.5,
        "perceived task demand" : 3
},{
"task 2" : "Avoiding vessels in the dark",
        "threat" : 1.5,
        "challenge" : 3,
        "task demand" : 4,
        "perceived task demand" : 3.32
},{
"task 3" : "Shadowing the target ship",
        "threat" : 1,
        "challenge" : 2.94,
        "task demand" : 5.67,
        "perceived task demand " : 2.37

},
```

Figure 5.2a Example of task file

```
"performance",
"constant" : -0.467,
"independent variables" : [

        {"name" : "HR-HRBaseline",
        "scalar" : -0.002
},{     "name" : "HRV-HRVBaseline",
        "scalar" : -0.651
},{     "name" : "threat",
        "scalar" : 0.122
},{     "name" : "challenge",
        "scalar" : 0.067
},{     "name" : "task demand",
        "scalar" : -0.059
},{     "name" : "perceived task
                  demand",
        "scalar" : 0.007

}]
```

Figure 5.2b Example of regression file.

$$\boldsymbol{performance\ prediction} = -0.467 + ((HR - baseline) * -0.002) + ((HRV - baseline) * -0.651) + (threat * 0.122) + (challenge * 0.067) + (task\ demand * -0.059) + (perceived\ task\ demand * 0.007)$$

The input values for the variables are then filled in with the values from task 1 in figure 5.2a. If a trainee using the system has a current heart rate of 85 beats per minute (bpm) with a baseline HR of 70 bpm and a heart rate variability of 0.80 with a baseline HRV of 0.65, the formula to predict performance is as follows:

$$\boldsymbol{performance\ prediction = -0.56} = -0.467 + \big((85 - 70) * -0.002\big) + \big((0.80 - 0.65) * -0.651\ \big) + (1 * 0.122) + (2.39 * 0.067) + (4.5 * -0.059) + (3 * 0.007)$$

5.3 FEEDBACK SYSTEM MODULES

The user interface of the COPE-FB system consists of two parts: one for the trainer and one for the trainee. The trainer part will be used by the trainer, or, as is the case in Chapter 6 and 7, by the experimenter. Figure 5.4 shows a screen shot of the trainer part.

5.3.1 TRAINER MODULE

For the COPE-FB system to become operational, the right settings need to be provided. A file containing the regression models needs to be selected (1) (see figure 5.2b). Every new scenario that will be used with the COPE-FB system needs regression model calibration, and will therefore result in different models. Next, the trainer selects the task files (2) (see figure 5.2a). Based on the data from the experiment at the Navy in Chapter 4, there are two files: one file contains input values for beginners and the other contains the input values for experts. Then, the heart rate device needs to be connected via Bluetooth (3). The trainer screen also shows the measured heart rate and heart rate variability. The trainer can than select which parts of the feedback information will be shown to the trainee (4). Different combinations of predicted performance feedback, physiological feedback, and predicted error-chance feedback can be selected for a training session. Selecting or deselecting three types of feedback results in a total of seven possible combinations. In Chapter 7, all these different combinations are tested on their effectiveness.

Next, the trainer had to set the feedback calculation interval (6). This interval determines how long the feedback is presented before new predictions are calculated.

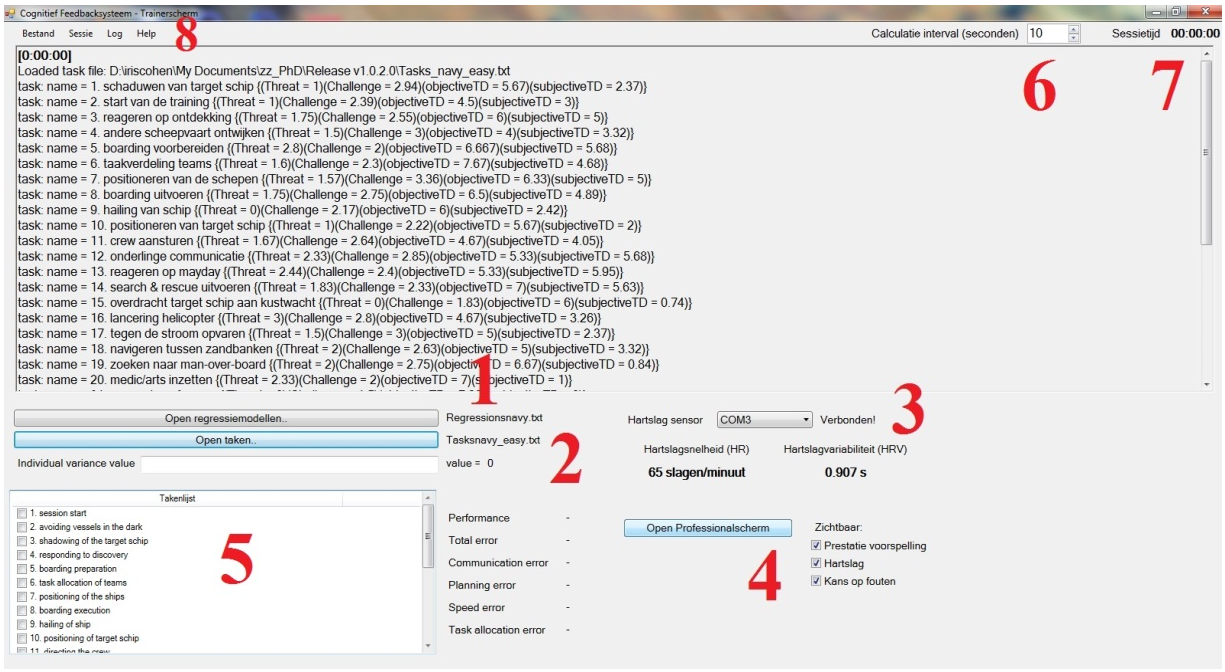


Figure 5.4 The trainer module of the COPE-FB system. This module controls the feedback system.

After all these settings are established, the trainer can start and stop the actual feedback session (8). The duration of the session is shown in the top right panel (7). Whenever the trainee switches to a different task, the trainer selects the task being performed (5). This triggers the feedback system to read the appropriate variable values from the task file.

5.3.2 TRAINEE MODULE

Figure 5.5 illustrates the trainee screen of the COPE-FB system. It shows the feedback types chosen by the trainer or experimenter in the trainer module. In the figure, all three types are selected and measured. Predicted performance feedback (1) shows the predicted performance on a scale of 0 to 10, physiological feedback (2) shows the current heart rate of the trainee on a scale from 30 to 130 bpm, and the predicted error-chance feedback (3) shows the chance on four types of errors between 0 and 10. Error-chances are predicted with logistic models that provide a chance value between 0 and 1. These values are converted to the scale 0 to 10.

5.3.3 ADDITIONAL OPTIONS

Apart from the options in the trainer and trainee modules, there are some additional options. First, the settings and the performance predictions can be saved manually (and automatically after closing the system) in a log file. As illustrated in figure 5.4, a

Figure 5.5 The COPE-FB system feedback. This is how the feedback is presented to the trainees.

dropdown menu for the log files is present at point 8 in the trainer module. Furthermore, a (Dutch) manual can be selected from the 'help' button at the top left panel in the trainer module (8). The help button also provides copyright information.

5.4 Conclusion

This chapter introduced the COPE-FB system. It explained the reasoning behind the chosen types of feedback and behind the commencement and duration, as well as the design of the feedback. The differences between the two modules (trainee and trainer) were explained and instructions were provided on how to set up the COPE-FB system for use.

The next chapters evaluate the effectiveness of the complete system on performance in a stressful simulated naval scenario (Chapter 6) and the different feedback types and combinations of feedback types during a stressful naval scenario in a low-fidelity simulator (Chapter 7).

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. The journal of the learning sciences, 4(2), 167-207.

Bouchard, S., Bernier, F., Boivin, E., Morin, B., & Robillard, G. (2012). Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. Plos one, 7(4).

Dörner, D., & Schaub, H. (1994). Errors in planning and decision-making and the nature of human information processing. Applied psychology: an international review, 43(4), 433-453.

Gohm, C. L., Baumann, M. R., & Sniezek, J. A. (2001). Peronality in extreme situations: thinking (or not) under acute stress. Journal of research in personality, 35, 388-399.

Gonzalez, C. (2005). Decision support for real-time, dynamic decision-making tasks. Organizational Behavior and Human Decision Processes, 96, 142–154.

Keinan, G., Friedland, N., & Ben-Porath, Y. (1987). Decision making under stress: scanning of alternatives under physical threat. Acta psychologica, 64, 219-228.

Kenealy, P. M. (1997). Mood state-dependent retrieval: The effects of induced mood on memory reconsidered. The Quarterly Journal of Experimental Psychology: Section A, 50(2), 290-317.

Lerch, F. J., & Harter, D. E. (2001). Cognitive support for real-time dynamic decision making. Information Systems Research, 12(1), 63-82.

Mendl, M. (1999). Performing under pressure: stress and cognitive function. Applied animal behaviour science, 65, 221-244.

Prinsloo, G. E., Derman, W. E., Lambert, M. I., & Rauch, H. G. L. (2013). The effect of a single session of short duration biofeedback-induced deep breathing on measures of heart rate variability during laboratory-induced cognitive stress: a pilot study. Applied psychophysiology and biofeedback 38, 81-90.

Prinsloo, G. E., Rauch, H., Lambert, M. I., Muench, F., Noakes, T. D., & Derman, W. E. (2011). The effect of short duration heart rate variability (HRV) biofeedback on cognitive performance during laboratory induced cognitive stress. Applied cognitive psychology, 25(5), 792-801.

Rock, I., & Palmer, S. (1990). Gestalt psychology. Sci Am, 263, 84-90.

Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153-189.

Sime, J.-A. (2007). Designing emergency response training: seven ways to reduce stress. Paper presented at the International Conference on Cognition and Exploratory Learning in Digital Age, Algarve, PT.

Wickens, C. D., Lee, J., Liu, Y., & Becker, S. G. (2004). An introduction to human factors engineering (second ed.). New Jersey: Pearson Education.

6. A FEEDBACK SYSTEM BASED ON THE COGNITIVE PERFORMANCE AND ERROR MODEL: EFFECTIVENESS DURING TRAINING IN A VIRTUAL NAVAL SETTING

ABSTRACT

To improve performance during a virtual training under stress, a feedback system is being developed based on the COgnitive Performance and Error (COPE) model. This feedback system provides physiological (heart rate) feedback, predicted performance feedback and predicted error-chance feedback to the trainees in real-time. This chapter investigates the effectiveness of the feedback system. Naval students (n = 36) performed a stressful scenario in teams of three persons in a virtual ship environment, while receiving feedback either in the first half or in the second half of the scenario. At the end, the students rated the performed task on their perception of appraisal (challenge or threat) and task demand, and an expert (trainer) rated the students' task performance. The comments made by the trainers referring to trainees' actions during the scenario were analysed and counted as an indication of errors made by the trainees. During the feedback conditions significantly less errors were made regarding planning (N = 12, Z = 1.994, p= .046) and speed of task execution (N = 12, Z= 2.60, p =. 009) compared to the no-feedback condition. The performance scores rated by the trainers did not differ between the feedback and no-feedback condition. Trainees' experiences with the system were analysed with the System Usability Scale (SUS) and open questions. The system received a SUS rating in the lowest 20th percentile of normative data reported in literature. Also trainees' general impressions of the feedback system were mixed. Thirteen trainees thought positive about the system, 2 were neutral, and 12 were more negative towards the system. Eleven participants did not respond to this question. Overall, it seems that the feedback system is beneficial for reducing the amount of errors made while training in a virtual environment, however, there is substantial room for improving its effects and usability.

Keywords: training under stress, virtual reality, feedback, performance, simulator, navy

Chapter submitted as:

Cohen, I., Brinkman, W.P., & Neerincx, M.A. (2015) A feedback system based on the cognitive performance and error model: Effectiveness during training in a virtual naval setting.

6.1 Introduction

Professionals working at crises or disaster scenarios, have a relatively high risk for encountering uncertain and unexpected situations that bring along high levels of stress and information-processing demands (Driskell & Johnston, 1998). Naval ship operators, for example, encounter situations that require them to make decisions that may have severe consequences. These decisions need to be made under high levels of stress caused by, for example, having to process a great amount of complex information in a short period of time. High levels of stress can cause a decrease in cognitive functions such as: attention, memory formation, and memory recall (Kleider, Parrott, & King, 2010; Mendl, 1999; Orasanu & Backer, 1996). These cognitive functions are needed to execute processes such as decision making (Mendl, 1999).

Training professionals to operate under stress is a way to prepare them for these eventualities. This chapter reports a study that examined whether providing real-time feedback in such training could improve trainees' performances. The feedback provided is based on the Cognitive Performance and Error (COPE) model that is capable to predict performance and chances of specific errors (Cohen, Brinkman, & Neerincx, 2012). The model describes how a task is perceived and how this affects individual's cognition which leads to (in)appropriate decisions or reactions. It includes cognitive variables (such as appraisal and task demand) and physiological variables (such as: heart rate and heart rate variability). The model was implemented in the COPE-feedback (COPE-FB) system that provided trainees with predicted performance outcome and chances of errors in addition to physiological feedback. The effect of the system was examined during training in a virtual environment at the Royal Netherlands Naval College (RNNC).

6.1.1. Training environments and digital (decision) support tools

An effective method to train for working under stress is learning by experience (Beach & Lipshitz, 1993; Cesta, Cortellessa, & Benedictis, 2014). Creating realistic and stressful training settings in which trainees can gain experience of both the task and the stress without the risks of real-life disaster scenarios, can be done with Virtual Reality (VR). VR is able to increase stress levels measured with physiological variables (e.g. heart rate, cortisol) and subjective measures (Bullinger et al., 2005; Busscher, Vlieger, Ling, & Brinkman, 2011). It also seems possible to control the stress levels created by VR (Brouwer, Neerincx, Kallen, van der Leer, & ten Brinke, 2011; Hartanto et al., 2014).

Results presented by Kim, Rosenthal, Zielinski, and Brady (2014) suggest that different types of VR elicit different kinds of emotional responses. They found that a head mounted display (HMD) and a CAVE system, i.e. a virtual environment projected on walls and floors or ceiling, resulted in similar increases in self-reported emotional arousal but dissimilar changes in emotional valence. They also found that skin conductance was higher in more sophisticated virtual environment such as the CAVE compared to the HMD. Thus, more realistic environments are more likely to induce

stress with stressful stimuli (see also, (Smets, Abbing, Neerincx, Lindenberg, & van Oostendorp, 2010)). A recent meta-analysis (Ling, Nefs, Morina, Heynderickx, & Brinkman, 2014) seems to confirm this for VR systems designed to elicit anxiety. Across 33 studies, it found a significant correlation between the level of presence created by VR environments and anxiety reported.

Experiencing stress in VR enhances professionals' performances in real stressful situations (McClernon, McCauley, O'Connor, & Warm, 2010). Adding instructions provides more advantages, especially for the training of novices (Kirschner, Sweller, & Clark, 2006). Hence, next to the VR stress training, other training tools are still needed to help the trainee learn to perform their task under stress.

When stress is induced in a virtual training setting, controlling the stress reactions and its negative effects on performances is the next step. Salas, Driskell, and Hughes (1996) mention techniques to control for the undesirable physiological and psychological effects of environmental demands such as: cognitive and physiological control techniques. Biofeedback is such a physiological control technique that is used to reduce physiological stress reactions (Sime, 2007, December). Biofeedback techniques provide the trainee with insight into their physiological stress reactions and provide the trainee with the opportunity to decrease these reactions. The effects of biofeedback on stress reduction and related task performance improvements have been shown (Bouchard, Bernier, Boivin, Morin, & Robillard, 2012; Prinsloo, Derman, Lambert, & Rauch, 2013), but the long term effects of biofeedback are debated (Raaijmakers et al., 2013, September). Next to stress reduction, the insight into one's own emotional state or stress level is said to leave cognitive resources open for performing the task at hand (Driskell & Johnston, 1998; Gohm, Baumann, & Sniezek, 2001)

Biofeedback techniques teach individuals to control their physiological reactions to stress, but not their cognitive reactions (Gohm et al., 2001; Keinan, Friedland, & Ben-Porath, 1987; Mendl, 1999). Therefore, cognitive techniques should be added to stress control techniques (Driskell & Johnston, 1998). Early digital support systems, or Intelligent Decision Aids (IDAs) consisted of computers that would calculated and suggested the best decision (Kontogiannis & Kossiavelou, 1999). These support systems, however, showed some problems. First, they rarely showed decision improvements (Cohen, 1993). Second, the individuals using the systems were ahead of the tool (Cohen, 1993). And third, preprogrammed scenario's are limited to the tool designers imagination (Reason, 1987). Nowadays, IDAs collaborate with their users to reach a decision in one of two ways: (1) they provide the individual with more information and thus lower the information uncertainty, or (2) they support a person's cognitive strategies.

6.1.2. *COPE-FB SYSTEM*

In Chapter 5, the COPE FeedBack system (COPE-FB) was explained. This system provides real-time feedback on predicted performance, predicted error-chances, and physiological feedback. The physiological (bio-) feedback helps to control physiological stress reactions (Sime, 2007, December). To control cognitive reactions to stress, predicted performance feedback is provided. Performance outcome feedback should be combined with more detailed feedback for best results (Lerch & Harter, 2001), and pointing out error tendencies decreases the chance on these errors (Dörner & Schaub, 1994). The COPE-FB system contained one linear regression for performance predictions, one logistic regression for error predictions, and four logistic regressions for the prediction of different error categories. The errors were categorized into communication, planning, speed, and task allocation errors. All models needed six input values and six coefficients (or parameter values) for their predictions. The physiological input values were derived from a heart rate measure device and the two appraisal and task demand values were derived from a database-file. For every task that was performed, prior observations were needed to determine the appraisal and task demand values. The different functions of the COPE-FB system will not be explained again in this chapter.

6.1.3. *HYPOTHESIS*

The study presented in this chapter investigates the effectiveness of the COPE-based feedback system presented in Chapter 5. The main hypothesis of this study states that trainees' performances during virtual training will be higher and the number of errors will be lower when immediate feedback is provided compared to when no feedback is provided.

This study also examined whether the predictors of the COPE model, i.e. appraisal, task demand and arousal, were affected by the availability of feedback. In other words, whether feedback led to changes in the predictors and consequently resulted in change of performance.

The study was also intended as an overall test for immediate feedback. If the study demonstrated an effect for immediate feedback it justified further detailed research into the effect of the various components of immediate feedback that was presented in this study as single entity, i.e. the combination of physiological feedback, performance prediction and error prediction.

6.2. METHODS

6.2.1. *EXPERIMENTAL DESIGN*

The predictive models that were implemented in the feedback system were based on a study in the RNNC bridge simulators (Chapter 4) (Cohen, Brinkman, & Neerincx, 2015). The same settings were used in this experiment; during six session, a stressful scenario was played in the bridge simulators in which six Naval students cooperated in two teams of three to complete the scenario. The two teams were divided over two simulators. The scenario was split in two halves; both teams performed one half of the scenario with feedback and one half without feedback. This study was approved by the ethical committee at TNO Soesterberg and Delft University of Technology.

6.2.2. PARTICIPANTS

Participants (n = 36) were students at the Royal Netherlands Naval College. There were 25 male and 11 female participants. They had a mean age of 25.3 years with a standard deviation of 2.34. All participants had the same amount of experience in the Navy and had a Bachelor degree or a similar level from previous education.

6.2.3. SIMULATORS AND SCENARIO

Two high-fidelity bridge simulators from the RNNC in Den Helder, The Netherlands were used. These simulators had 360 degrees of view and replicated the bridge of real Naval ships. One of the simulators is shown in figure 6.1. The scenario that was performed was also used for establishing the predictive models that are used in this experiment (Cohen et al., 2015). The story took place at the North Sea, which was familiar territory for the participants. The scenario started with two Naval warships shadowing a ship that was suspected of smuggling refugees. This ship discovered that it was being followed, which meant they were likely to 'destroy evidence'. In other words: throwing refugees overboard. The participants then needed to board the smuggling ship. Before the ship could be boarded, several actions needed to be done. When the boarding was being executed, a Search-And-Rescue (SAR) call came in on the radio. Now, the two Naval ships needed to decide to follow the distress call or not. When they did, they needed to stop the boarding action and to hand it over to another virtual (coast guard) ship. When the SAR was being executed, several actions need to be done. All teams participated in the scenario for approximately 130 minutes. Not every session ended with the same number of tasks performed. This depended on how fast the team was working through all the previous events in the scenario.

In table 6.1, the episodes, goals and actions of the tasks as suggested by Ozel (2001) are explained. Six main episodes can be identified: (1) shadowing the target ship, (2) avoiding other vessels (this goal stays a goal during the whole experiment), (3) preparing for boarding, (4) executing the boarding, (5) reacting to a search-and-rescue (SAR) call and (6) executing a SAR. Within these main episodes, different actions can be identified, indicated in the third column of table 1.

Figure 6.1 Photograph of students training in the main bridge simulator.

6.2.4. MEASUREMENTS

TASK CHARACTERISTICS, APPRAISAL AND TASK DEMAND

The task related values for appraisal (threat and challenge) and task demand (perceived and estimated) had been measured in a previous study (Cohen et al., 2015). The average scores on these variables were now taken as the input values to create a real-time input.

Table 6.1. Actions that need to be executed in different stages of the scenario.

Episode/goal	Time in scenario	Stressful events: actions for episode goal
1) Start of the training	Start - ± 25 min	
2) Avoiding other vessels in the dark	During entire scenario	2a. Searching for suspicious vessel 2b. Intercepting suspicious vessel
3) Preparing to board target ship	± 25 min - ± 90 min	3a. Deciding what team does what 3b. Positioning of the ships
4) Executing combined boarding	± 35 min - ± 90 min	4a. Hailing the suspicious vessel 4b. Positioning the suspicious vessel 4c. Directing the crew 4d. Mutual communication
5) Reacting on incoming Mayday		5a. Transfer suspicious vessel to coastguard
6) Executing Search and Rescue	± 90 min – end	6a. Launch helicopter 6b. Gearing up against flow 6c. Navigate between sandbars 6d. Searching for 'man-over-board' 6e. Deploying the medic 6f. Carrying away injured

Additionally, the appraisal and task demand were also rated during this study to investigate the effects of feedback on appraisal and task demand. Both were rated on a 10-point scale, per task. For appraisal, this scale ran from 'task perceived as a challenge' (1) to 'task perceived as a threat' (10). The task demand scale ran from very low (1) to very high (10) task demand.

Performance rated by experts

The trainers that were present in both simulators rated task performance for every participant in the simulator, after the training sessions. They received a task list as in table 1, and rated the task performances on a 10-point scale. The scale ran from 1 (very low performance) to 10 (excellent performance).

Errors

Two Sony HDR-CX300E cameras were used to record the activities in the simulators. As in the previous study, the video images were used to determine what kind of comments were given by the trainers (Cohen et al., 2015). These comments acted as error indicators.

System Usability Scale (SUS)

A Dutch version of the System Usability Scale was used to measure the overall usability (Brooke, 1996), divided into learnability and usability (Lewis & Sauro, 2009), of the feedback system. One of the SUS questions was left out of the questionnaire since it was inappropriate for this feedback system. This was question 5: "I found the various functions in this system were well integrated".

Open questions

Next to the SUS, opens questions were asked to get information about the trainees' perception of the feedback system. These questions gave insight into the trainees' willingness to work with this feedback system in future training sessions. The following questions were asked:

- What is your general impression of the feedback system?
- How often and at what moments did you pay attention to the system?
- If you saw the feedback change, did you understand what it meant?
- What did you do when you saw performance/heart rate/errors change?
- Did you think other persons in the simulator saw your feedback? And did you find that unpleasant?
- Do you think that the presented feedback affected your performance/errors during the scenario?

- Would you like to use (a similar) feedback system more often during training?

6.2.5. PROCEDURE

For every session, the same procedure was followed. The participants received a simulator training session as part of their military education. The sessions used in this experiment was mainly focussing on crew resource management during stressful situations. Before the training session started, the trainers provided a briefing about the training session and the situation in which they would enrol when the scenario starts. Next, the experimenter informed the participants about this experiment, and handed out participant information forms. When they agreed to participate, they signed a consent form and filled a short questionnaire with participant characteristics (age, gender, experience in the simulator and number of simulator and current role). None of the students refused participation. The feedback system was also briefly explained.

The participants formed two teams based on their own preference and collaboration experience and were assigned to one of the simulators. Each team had a participant fulfilling the role of an officer of the watch, a navigation officer and a steersman. The officer of the watch had an overall leading role and made sure the other participants performed their tasks. The navigation officer, besides navigating the ship, helped the officer of the watch with communication tasks, for example calling the target ship to perform an identification check. The steersman mostly steered and maneuvered the ship. In the simulator the heart rate devices were put on as shown in figure 6.2. While heart rate and heart rate variability baseline measurement took place, the scenario was loaded into the simulator and an extra explanation form about the feedback system was handed out to the participants. This form explained the different bar graphs in the feedback screen, and how they should be interpreted. The trainee feedback screens were displayed on a separate monitor, closely located to the participants working area. The brightness and contrast of these displays was turned to the lowest settings to prevent bright light sources in a dark simulator.

The experiment had a split-plot design. Participants experienced the feedback condition as well as the no-feedback condition. To counterbalance the order of the feedback condition, the feedback was provided in one simulator until a 15 minute break halfway into the scenario. After this break the feedback set-up was turned off in the initial simulator and turned on in the other simulator. The participants stayed in the simulator in which they started. The order of the feedback versus no feedback condition was constant for both simulators, meaning simulator 1 always started with feedback and ended the scenario without feedback and vice versa for the second simulator.

Figure 6.2 The HxM Zephyr Bluetooth belt and its placement on the participants' chest.

6.2.6. DATA PREPARATION AND DATA ANALYSES

Missing data were caused by technical problems with the COPE-FB system, unfinished questionnaires either by the trainers or the trainees, and missing video data. When the COPE-FB system was not properly working for one participant, this participant usually could not fill in the System Usability Scale and other usability questionnaire at the end of the session. Few participants failed to fill in the backside of the questionnaires. Data was analysed in pairs (one case with date from the feedback condition and the no-feedback condition), which meant that when data was missing from one condition, the entire case was excluded from the analyses. There were performance scores for 24 pairs, appraisal data for 22 pairs, task demand data for 21 pairs and error data for 12 pairs.

Trainers filled in performance questionnaires for 24 trainees. In the cases where ratings were obtained from two trainers, the ratings were averaged to create a single performance score per trainee per task. The appraisal and task demand questionnaire was completed by 22 trainees. For both the appraisal and task demand a single average scores was calculated per participant for the actions in the first half of the simulator sessions (task 1-15) and for the tasks in the second half of the sessions (task 16-21).

The video analyses resulted in a list of 103 comments (in Dutch) made by the trainers during the simulator sessions. These comments were categorized into four error categories: planning, communication, speed and task allocation. Three native-Dutch coders, unaffiliated with the Navy, categorized all the comments into one of the categories. The main coder (first author of this chapter) first categorized all the comments. Next, two coders with the same native language categorized the comments as well. Coder 1 and coder 2 had inter-rater correlations of Cohen's κ between 0.56 and 0.86 and between coder 1 and coder 3 the Cohen's κ was between 0.32 and 0.78. The cases where the two coders deviated from the categorization of the main coder were discussed and categorized again by the two coders. This process was repeated and led to the final inter-rater agreement scores measured in Cohen's κ's, as shown in table 6.2, for all categories between the main coder and coder 1 and coder 2. If two or more coders agreed on a comment, this categorization was used for the final categorization of the errors. The different errors were counted per participant and per session half as was done with the performance and task characteristics.

Table 6.2. Agreement of error categorisation between coder 1 and coder 2 and 3. Measured in Cohen's κ's.

	Coder 1 vs Coder 2	Coder 1 vs Coder 3
Communication	0.98	0.95
Planning	0.86	0.87
Speed	1.00	0.90
Task allocation	0.88	0.88
Others	0.95	0.95

Nineteen participants filled in the System Usability Scale after finishing the experimental sessions. The other participants did not finish the questionnaires and therefore did not complete the SUS. Normally, the SUS has a range from 0-100 but in this experiment, one question was left out of the SUS. To keep the range of the SUS at 0-100 and not 0-90, an alteration was made in the scoring. Normally when the even and uneven questions are scored, they are summed and multiplied by 2.5. In this experiment, the sum was however multiplied with 2.778 to keep the total maximum score at 100.

In the open questions, participants were asked for the general impression of the system. The answers consisted of comments on the system such as: "very interesting" or "too much information". These comments were rated on their positivity by the same three coders who had coded the trainee's errors. They rated the remarks as either, negative, neutral, or positive. When two or more coders agreed, this rating was used, no cases occurred where all three raters disagreed. The ratings of the participants' comments had an inter-rater agreement ranging between Cohen's κ of .77 and .78.

The data was analysed using non-parametric statistical tests in SPSS 20, such as Wilcoxon Signed Ranks test for comparing COPE variables and performances, and with binominal test to test whether remarks towards the feedback system trainees made were dominantly positively or negatively.

6.3. RESULTS

6.3.1. EFFECT OF FEEDBACK

The first analysis examined trainees' performance scores as rated by the trainers. A Wilcoxon's signed ranked test found no significant difference (N = 24, Z = 0.31, p = 0.75) in the performance rating between the without feedback condition (M = 6.25, SD = 0.96) and the condition with feedback (M = 6.04, SD = 1.01).

The second analysis examined the number of errors made by the trainees. Table 6.3 shows the descriptive statistics for all the error categories. The number of errors made was low in general which makes the reduction low as well (total max = 15). A Wilcoxon's signed ranked test on the total number of errors made during the with- and without feedback conditions, found a trend towards less errors when feedback is provided, but no significant difference (N = 12, Z = 1.73, p = 0.084). This analysis was repeated, this time excluding the errors classified as 'other', as they were not included in feedback given to trainees. This analysis also showed a trend towards less errors when feedback was provided but no difference was found (N = 12, Z =-1.794, p =0.073). Further detailed analyses focussed on the different error categories. For two error categories a significant difference was found in number of errors made during the without feedback and the feedback condition. The number of planning errors when no feedback was provided (M = 1.08, SD = 1.31) was significantly higher (N = 12, Z = 1.994, p= .046) than the number of planning errors made when feedback was provided (M = 0.40, SD = 0.76). A difference was also found in number of errors referring to speed. When no feedback was provided the number of speed errors (M = 1.67, SD = 1.07) was significantly higher (N = 12, Z= 2.60, p =. 009) than when feedback was provided (M = 0.92, SD = 0.86).

The last comparison examined the appraisal and task demand scores during the different conditions. For appraisal, a Wilcoxon Signed Ranks test found no significant difference (N = 22, Z =-0.47, p = 0.64) between the condition without feedback (M = 3.65,

Table 6.3. Descriptive statistics of the number of errors occurring during feedback and no-feedback conditions (N = 12 pairs).

	M	SD	Median	Mode	Min.	Max.
Feedback						
Total	3.46	2.58	3	3	0	10
Communication	1.00	1.32	1	0	0	5
Planning	0.40	0.76	0	0	0	3
Speed	0.92	0.86	1	1	0	3
Task allocation	0.33	0.87	0	0	0	4
Other	1.12	1.17	1	0	0	4
No feedback						
Total	6.75	4.39	5.5	5	1	15
Communication	2.00	1.60	1.5	1	0	5
Planning	1.08	1.31	0.5	0	0	3
Speed	1.67	1.07	1	1	1	4
Task allocation	0.25	0.45	0	0	0	1
Other	1.83	1.34	1.5	1	0	4

SD = 1.51) and with feedback (M = 4.16, SD = 1.70). A similar analysis on task demand, also failed to find a significant difference (N=21, Z=-0.122, p = 0.90) between without feedback (M = 5.72, SD = 1.21) and with feedback (M = 5.92, SD = 1.71) condition.

6.3.2. USABILITY AND TRAINEE REMARKS OF THE COPE-FB SYSTEM

The COPE-FB system received a total mean SUS score of 50.2, with a minimum of 25 and a maximum of 72.23. According to Sauro (2011), this score would be in the lowest 20th percentile compared to SUS data from over 5000 users over 500 evaluations.

Next to the SUS, the participants also filled in a questionnaire with open questions about their experience with the feedback system. Twenty-two participants filled in these open questions. When asked how often they watched the feedback screen, only ten participants replied. They answered between 0 times and 11 times with an average of 5.9 times. Participants mainly looked at the feedback screen when they were not busy (n = 9) or at random times (n =4). Other moments that were used to look at the feedback were (n = 1 for all following options) 'when I was busy', 'when I finished a specific task', 'continuously', or 'at the beginning of the scenario'. An open question asked about the participants' general impression of the COPE-FB system. These impressions were rated as positive, neutral or negative. The general impression was 13 times positive, twice neutral and 12 times negative. Fifteen different statements were given by 27

Table 6.4. Frequencies and positivity ratings of the general impressions as stated by the participants. "+" = positive, "-" = negative "0" = neutral.

General impression	n	positive/negative
Overall positive	5	+
Interesting	5	+
Overall negative	3	-
Too much information	3	-
Doubts about usefulness	2	-
Professional	1	+
Part relevant, looked at it every now and then	1	+
Did not work (HxM connectivity)	1	-
To me it was not clear what the results meant	1	-
I did not watch it, no impression	1	0
Nice, but difficult to check since we have a primary task to perform.	1	0
No time to deal with it	1	-
Results lag behind, extreme values	1	-
I thought it was useful to keep myself calm	1	+

participants when asked about their general impressions. These are listed in table 6.4, with a positivity rating in the third column. A binominal test showed that there is no preference towards either positive or negative impressions of the system when the two neutral impressions were left out (n =27, expected probability 0.5, p. > 0.99).

It is important that the users of the COPE-FB system understand what the feedback means, or what they can do to improve their performance. The questionnaire showed that seven participants indicated that they understood why feedback was changing and seven participants indicated that they did not understand why feedback was changing.

Participants were also asked what they did when the feedback changed (table 6.5). When performance outcome and error chance feedback changed most participants (3 and 5 respectively) reported that they did nothing. When physiological feedback changed, most participants (n=6) reported a specific strategy that they applied (breath calmly).

Table 6.5. What participants did when specific feedback changed.

Performance feedback	n	%	Error feedback	n		Physiological feedback	n	%
Nothing	3	37.5%	Nothing	5	50%	Breath calmly	6	42.9%
Breath calmly	2	25.0%	Anticipate / keep in mind what to improve	2	20%	Nothing	5	35.7%
Keep up the work or try harder	1	12.5%	Unclear	1	10%	Stay/become calm	1	7.1%
Kept in mind while performing the tasks	1	12.5%	Tried to create calmness	1	10%	'take a step back'	1	7.1%
Not much	1	12.5%	Not much	1	10%	Not much	1	7.1%

Table 6.6. Did feedback influence your performance and the number of errors you made?

Influence on performance	*n*	%	Influence on errors	*n*	%
No	15	62.5%	No	16	72.7%
Yes	3	12.5%	Yes	1	4.6%
Don't know	1	4.2%	Both	1	4.6%
Did not see it often enough (no)	2	8.3%	No time for interpretation (no)	2	9.1%
At the beginning (yes)	2	8.3%	Did not see if often enough (no)	1	4.6%
Not for my role (no)	1	4.2%	Not for my role (no)	1	4.6%

When asked if they thought if other people present in the simulator looked at their feedback screen, seven people said yes and eight people said no. When asked if they thought this was unpleasant, two people said yes, ten people said no. One participant elaborated and said that it was no big deal because the people in the simulator were his/her classmates, but one participant indicated it was not pleasant if subordinates could see your heart rate go up.

Participants also rated if they thought the system influenced their performance level and number of errors made (table 6.6). The majority (62.5% and 72.7%) said they did not believe performance or errors were influenced by the feedback system. When asked if they would like to use a similar system again during training, 11 trainees said "yes" and five said "no". A few participants that responded with "yes" also gave suggestions such as: focus the system more on heart rate, make the system more visible in the simulator and apply only for certain roles in the simulator. Participants that said "no", mentioned that they did not see added value and one participant said he/she would rather have had human feedback.

6.4. DISCUSSION

6.4.1. CONCLUSIONS

The main conclusion that can be drawn from the findings is that immediate feedback from the COPE-FB system has an added value for trainees. Although analysis did not found a difference in performance ratings by trainers between the with- and without feedback condition, the findings did show a reduction in the number of planning errors and errors regarding the speed of the task execution. The main hypothesis of the chapter is therefore partly supported; providing trainees with immediate feedback increased parts of their performance. It should be noted that the number of errors were rather low in general and, consequently, the reduction in errors small. Future research should try to use a more sensitive objective performance measure.

Since the analyses did not show a change in appraisal and task demand between the with- and without feedback conditions, it is not sure whether these variables could explain the performance improvement. This is interesting as both variables are put forward in the COPE model as performance predictors. The other predictor, arousal, was not examined in this manner and could therefore be responsible for the performance improvement. Since these variables are not responsible for the improvements, as suggested by the COPE model, other factors could have played a role as well. Getting feedback during training might draw trainees' attention to the likelihood of making an error and therefore stimulating them to avoid these errors instead of influencing the COPE-variables.

The System Usability Scale does not function as a diagnostic tool but when the score is low, it is an indication that there is room for user interface improvement (Sauro, 2011). The mean SUS score for the COPE-FB system was 50.2. Bangor, Kortum, and Miller (2009) added a 7-point acceptability scale to the SUS. A SUS score of 50.2 relates to the middle score of "OK" on this acceptability scale (Bangor et al., 2009). This acceptability scale, however, was based on different types of systems (e.g. websites, cell phones, TVs, graphical user interfaces) than the COPE-FB system. The qualitative data provides some clues as to what areas to improve. The analysis of trainees' remarks is roughly equally divided among trainees regarding their attitude towards the feedback system. An insightful observation was that a large group of trainees did not belief that changes in the feedback indicator led to changes in their behaviour. As a significant reduction was found in number of errors, it suggests that trainees might not be aware of this. Also they might not be aware of an explicitly strategy they applied or should apply as a response to changes in feedback indicators. Still an exception was physiological feedback as a large group of the trainees indicated breathing calmly as strategy for increased heart-rate. This suggests that when using a feedback system it cannot be assumed that trainees understand why indicators change, and what they should do when indicators change. Future work should therefore examine whether addressing this gap in knowledge prior to training could enhance acceptance and performance.

6.4.2. Limitations

Some limitations that mitigate the findings related to the decreased in number of errors should be noted. First of all, the error data suffered a loss of data due to missing video data. The errors were derived from the trainer comments during the training. The trainers knew and saw that there were feedback screens in the simulator which could have biased the trainers and make them comment less on the trainees. The displays that showed the feedback were located at fixed locations, which is another limitation since the trainees walked around in the simulator. The feedback was therefore not always in the visual field of the trainees. In this paper, participants indicated how often they looked at the feedback screen. This is a subjective measure and in further studies this could be measured objectively with an eye tracking device. Such measures will give more accurate data, and can also show what part of the feedback the participants are focussing on. Another limitation that could have distorted the analyses is that the trainees performed in teams. The COPE-FB system provides feedback aimed at individuals and did not take into account the effect of other team members. Furthermore, the predictive feedback models did not reach a 100% accuracy (Cohen et al., 2015). Since the predictions are based on subjective trainer ratings and comments, an accuracy of 100% might not be possible.

The qualitative data showed that the trainees did not completely comprehend the feedback, which indicates an inadequate amount of explanation of the feedback system. Instruction time was limited since the experiment took place during a routine Naval

training. The trainees did receive a written explanation but it is not clear if they read the whole explanation. Future use of the COPE-FB system might benefit from a more detailed explanation and maybe a short training with the system.

Another note that should be made with regard to the effect on errors is that this effect could be of short term. Participants see the feedback and pay more attention to these categories of errors. Further studies are needed to find out if the feedback improves the acquisition of skills and has long-lasting effects on the performance.

6.4.3. IMPLICATIONS

In spite of the limitations, the results of this experiment show promising effects of the COPE-FB system. Some significant results indicate that the COPE-FB system is at least partly effective and the qualitative data shows that participants are interested in receiving feedback. Other qualitative data indicate areas for improvement (e.g. focus the system on the physiological feedback, make the screens mores visible, adapt the feedback to the roles in the scenario) that might lead to a more effective COPE-FB system.

Some limitations could be overcome with alterations in the design of the system. If the system is displayed in a hand-held device, the problem with the lack of visibility will be solved. Trainees indicated that they watched the feedback when they were not busy while these moments might be the ones that need feedback. Adding a warning signal when the error chances reach a certain threshold could focus attention to the feedback for at least a short period of time. If the feedback is implemented into a hand-held device attached to the trainees' wrist, a tactile warning signal is preferred since it will not distract other trainees. Using a device like this will also solve potential privacy issues (Harbers, Aydogan, Jonker, & Neerincx, 2014, May 5-9). The possibility of using a Head-Mounted Display to present the feedback can also be investigated. Trainees' performance and stress levels will no longer be visible for other teammates or subordinates.

The points raised in this discussion suggest that some further studies and alterations in the system are necessary before the COPE-FB system is adopted into scenario-based training sessions. After that, another question might be of interest. Would this tool be beneficial for real-life scenarios, outside the training simulator? It is too soon to say that this is an option, since the study conducted here involved trainees with a lower experience level. More experienced professionals might have different thoughts or preferences towards feedback systems. Nonetheless, when performance improves due to the COPE-FB system, it is worth investigating if this system should also be used outside the training location.

ACKNOWLEDGEMENT

The work presented in this chapter is supported by the Dutch FES program: Brain and Cognition: Societal Innovation (project no. 056-22-010). We would like to thank the KIM in Den Helder, The Netherlands, especially the trainers at the bridge simulator for their help with running the experiment and scheduling the participants. Furthermore we thank the coders Corine Horsch and Myrthe Tielman for their time and effort.

REFERENCES

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of usability studies, 4, 114-123.

Beach, L. R., & Lipshitz, R. (1993). Why classical decision theory is an inappropriate standard for evaluating and aiding most human decision making. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), Decision making in action: Models and methods (pp. 3-20). Norwood, New Jersey: Ablex publishing corporation.

Bouchard, S., Bernier, F., Boivin, E., Morin, B., & Robillard, G. (2012). Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. Plos one, 7(4), e36169.

Brooke, J. (1996). SUS: a quick and dirty usability scale. In P. W. Jordan, B. Thomas, I. L. McClelland & B. Weerdmeester (Eds.), Usability evaluation in industry (pp. 189): CRC Press.

Brouwer, A.-M., Neerincx, M. A., Kallen, V., van der Leer, L., & ten Brinke, M. (2011). EEG alpha asymmetry, heart rate variability and cortisol in response to virtual reality induced stress. Journal of Cybertherapy & Rehabilitation, 4, 21-34.

Bullinger, A. H., Hemmeter, U. M., Stefani, O., Angehrn, I., Mueller-Spahn, F., Bekiaris, E., . . . Mager, R. (2005). Stimulation of cortisol during mental task performance in a provocative virtual environment. Applied psychophysiology and biofeedback, 30, 205-216.

Busscher, B., Vlieger, D. d., Ling, Y., & Brinkman, W.-P. (2011). Physiological measures and selfreport to evaluate neutral virtual reality worlds. Journal of Cybertherapy & Rehabilitation, 4, 15-25.

Cesta, A., Cortellessa, G., & Benedictis, R. D. (2014). Training for crisis decision making - An approach based on plan adaptation. Knowledgde-based systems, 58, 98-112.

Cohen, I., Brinkman, W.-P., & Neerincx, M. A. (2012). Assembling a synthetic emotion mediator for quick decision making during acute stress. Paper presented at the Proceedings of the 2012 European Conference on Cognitive Ergonomics, Edinburgh.

Cohen, I., Brinkman, W.-P., & Neerincx, M. A. (2015). Modelling environmental and cognitive factors to predict performance in a stressful training scenario on a naval ship simulator. Cognition, Technology & Work, 2015, 1-17.

Cohen, M. S. (1993). The bottom line: naturalistic decision aiding. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), Decision making in action: Models and methods (pp. 3-20). Norwood, New Jersey: Ablex publishing corporation.

Dörner, D., & Schaub, H. (1994). Errors in planning and decision-making and the nature of human information processing. Applied psychology: An international review, 43, 433-453.

Driskell, J. E., & Johnston, J. H. (1998). Stress exposure training. In J. A. Cannon-Browers & E. Salas (Eds.), Making decision under stress: Implications for individual and team training (Vol. 3, pp. 191-218). Washington, DC: American Psychological Association.

Gohm, C. L., Baumann, M. R., & Sniezek, J. A. (2001). Peronality in extreme situations: thinking (or not) under acute stress. Journal of research in personality, 35, 388-399.

Harbers, M., Aydogan, R., Jonker, C. M., & Neerincx, M. A. (2014, May 5-9). Sharing information in teams: giving up privacy or compromising on team performance? Paper presented at the Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, Paris, France.

Hartanto, D., Kampmann, I. L., Morina, N., Emmelkamp, P. G., Neerincx, M. A., & Brinkman, W.-P. (2014). Controlling Social Stress in Virtual Reality Environments. Plos one, 9(3), e92804.

Keinan, G., Friedland, N., & Ben-Porath, Y. (1987). Decision making under stress: scanning of alternatives under physical threat. Acta psychologica, 64, 219-228.

Kim, K., Rosenthal, M. Z., Zielinski, D. J., & Brady, R. (2014). Effects of virtual environment platforms on emotional responses. Computer methods and programs in biomedicine, 113(3), 882-893.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational psychologist, 41, 75-86.

Kleider, H. M., Parrott, D. J., & King, T. Z. (2010). Shooting behaviour: how working memory and negative emotionaliy influence police officer shoot decisions. Applied cognitive psychology, 24, 707-717.

Kontogiannis, T., & Kossiavelou, Z. (1999). Stress and team performance: principles and challenges for intelligent decision aids. Safety science, 33, 103-128.

Lerch, F. J., & Harter, D. E. (2001). Cognitive support for real-time dynamic decision making. Information Systems Research, 12, 63-82.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. Heidelberg: Springer Berlin.

Ling, Y., Nefs, H. T., Morina, N., Heynderickx, I., & Brinkman, W.-P. (2014). A meta-analysis on the relationship between self-reported presence and anxiety in virtual reality exposure therapy for anxiety disorders. Plos one, 9(5), e96144.

McClernon, C. K., McCauley, M. E., O'Connor, P. E., & Warm, J. S. (2010). Stress Training Enhances Pilot Performance During a Stressful Flying Task. Human Factors, 53, 207-218.

Mendl, M. (1999). Performing under pressure: stress and cognitive function. Applied animal behaviour science, 65, 221-244.

Orasanu, J. M., & Backer, P. (1996). Stress and military performance. In J. E. Driskell & E. Salas (Eds.), Stress and human performance (pp. 89-125). Mahwas, NJ: Lawrence Erlbaum Associates, Inc.

Ozel, F. (2001). Time pressure and stress as a factor during emergency egress. Safety science, 38, 95-107.

Prinsloo, G. E., Derman, W. E., Lambert, M. I., & Rauch, H. G. L. (2013). The effect of a single session of short duration biofeedback-induced deep breathing on measures of heart rate variability during laboratory-induced cognitive stress: a pilot study. Applied psychophysiology and biofeedback 38, 81-90.

Raaijmakers, S. F., Steel, F. W., Goede, M. d., Wouwe, N. C. v., Erp, J. B. F. v., & Brouwer, A.-M. (2013, September). Heart rate variability and skin conductance biofeedback: A triple-blind randomized controlled study. Paper presented at the Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva.

Reason, J. (1987). Cognitive aids in process environments: prostheses or tools? International journal of man-machine studies, 27, 463-470.

Salas, E., Driskell, J. E., & Hughes, S. (1996). Introduction: the study of stress and human performance. In J. E. Driskell & E. Salas (Eds.), Stress and Human Performance (pp. 1-45). Hillsdale, NJ: Erlbaum.

Sauro, J. (2011). Measuring usability with the System Usability Scale (SUS) Retrieved December 2014

Sime, J.-A. (2007, December). Designing emergency response training: seven ways to reduce stress. Paper presented at the International Conference on Cognition and Exploratory Learning in Digital Age, Algarve, Portugal.

Smets, N., Abbing, M., Neerincx, M., Lindenberg, J., & van Oostendorp, H. (2010). Game-based versus storyboard-based evaluations of crew support prototypes for long duration missions. Acta Astronautica, 66(5), 810-820.

7. REAL-TIME FEEDBACK ON PHYSIOLOGICAL, PREDICTED PERFORMANCE AND PREDICTED ERROR-CHANCES FOR PERFORMING IN HIGH-DEMANDING WORK CONDITIONS

ABSTRACT

Experiencing stress during training is a way to prepare professionals for acting in real-life crises. With the help of feedback tools, these professionals can train to recognize and overcome negative effects of stress on task performances. This paper reports a study that empirically examined the effect of such a feedback system. The system, based on the COgnitive Performance and Error (COPE) model, provided physiological, predicted performance and predicted error-chance feedback. It was calibrated and tested in two experiments. The first experiment focussed on creating stressful scenarios and establishing the parameters for the predictive models for the feedback system. In a virtual ship environment, participants (n = 9) had to extinguish fires. Stress was induced by altering time pressure, information uncertainty and consequences of performance. During the task execution, COPE variables were measured to establish predictive models. For every 10 seconds, a model predicted performance with a Spearman's rho = 0.12 correlation towards the observed performance values, and models were established that predict the chances of specific errors with area under the curve values between 0.73 and 0.8. In the second experiment a new group of participants (n = 29) carried out the same task while receiving eight types of feedback in a counterbalanced order. The feedback consisted of all possible combinations of the three feedback types of the COPE based system. Performance scores improved when feedback was provided during the task compared to not receiving feedback. The number of errors made did not improve. No further effects of different types of feedback on performance and errors were found. However, the usability score of the system with physiological feedback was significantly higher than a system without feedback. Especially when physiological feedback was not combined with error feedback. The latter result was confirmed by participants reporting that they understood the physiological feedback but got distracted when too much feedback was provided. The findings in this paper show effects of feedback on performances and usability. To improve the effectiveness of the feedback system it is suggested to provide more in-depth tutorial sessions. Design changes are also suggested that would make the COPE FeedBack system more effective in improving performances.

Keywords: stress, virtual training, decision tools, cognitive errors, task performance

Chapter submitted as:

Cohen, I., Brinkman, W.P.. & Neerincx, M.A. (2015) Real-time feedback on physiological, predicted performance and predicted error-chances for performing in high-demanding work condition. *International Journal of Human-Computer Studies*.

7.1. INTRODUCTION

For professionals to be ready to work under stressful circumstances, such as crises, combat or disaster scenarios, they need proper training. An effective way to train the skills for decision making under stress, is learning by experience (Andresen, Boud, & Cohen, 2001). This means that these skills are best learned while experiencing stress conditions that match the experiences of the real practice (Beach & Lipshitz, 1993; Cesta et al., 2014). In Virtual Reality (VR), scenario-based training environments can be created that provide the realistic stressful situations (Peeters et al., 2014). VR seems to be able to elicit physiological stress responses in individuals (Busscher et al., 2011; Hartanto et al., 2014) but leaves out the risk of real-life crisis and disasters.

Experiencing stress in VR enhances professionals' performances in real stressful situations (McClernon, McCauley, O'Connor, & Warm, 2010). Adding instructions to such training would provide more advantages, especially for the training of novices (Kirschner, Sweller, & Clark, 2006). Hence, next to the VR training, other training tools are needed to help the trainee learn to perform tasks under stress.

This chapter shows that assisted learning can also improve performance during stressful situations. The assisted learning type that is argued for, is real-time feedback provided in a simulation-based training. The motivation behind this work lies in the benefits computers could bring in the acquisition of knowledge about the cognitive and affective processes and their outcomes in simulated stressful situations.

One way of assisting trainees in VR would be to incorporate decision support systems into the learning environment. These interactive systems can be divided in two categories: cognitive prostheses and cognitive tools (Wickens et al., 2004). The first category includes systems that replace cognitive decision-making processes for the users. The users merely provide the system with data and information that is needed for to produce a decision. Cognitive tools on the other hand, are designed to provide support to the decision makers instead of replacing them.

Cognitive prostheses work well in a clearly defined decision making situation, but they do not seem to work successfully in uncertain situations since they can only make decisions on pre-programmed situations (Reason, 1987). Furthermore, these systems rarely show improvement in the decisions being made since human decision processes are often ahead of the system (Cohen, 1993). People are also reluctant to accept being subordinate to a system (Gordon, 1988; Kontogiannis & Kossiavelou, 1999; Wickens et al., 2004). Cognitive prostheses also change the nature of a task from a decision task to learning to understand the system. Cognitive tools, however, might be more appropriate for use in training settings as they support the user in learning a skill. In real-life settings, a cognitive tool can still help professionals to be more aware of negative effects of stress. Another reason to prefer cognitive tools over cognitive prostheses in uncertain environments is that professionals do not seem to pick a decisions after considering several alternatives (Klein et al., 1986). Therefore, a support tool should not support this type of decision making.

Effective support tools for uncertain situations should focus on skill or knowledge enhancement, or control. Biofeedback methods, for example, have been used for personal stress control. These methods provide individuals with an insight into their physiological reactions to stress. Trainees try to reduce these reactions and over time, they learn to control the physiological reactions to stress. Being more aware of one's emotional state is said to leave more cognitive resources for the task (Driskell & Johnston, 2006; Gohm et al., 2001). Some studies demonstrated biofeedback's ability to reduce stress, and consequently improve performance (Bouchard et al., 2012; Prinsloo et al., 2013), but these findings may be biased due to un-blind trials (Raaijmakers et al., 2013).

Current support tools generally offer three types of feedback: (1) outcome feedback states the current performance of a task, (2) cognitive feedback explains how to perform the task and, (3) feed-forward helps the user to anticipate on different decision options. When outcome performance is provided on its own, it does not seem to be effective in increasing task performance (Gonzalez, 2005; Lerch & Harter, 2001). It might still put the trainee in an unguided learning situation. Combining it with feed-forward feedback on the other hand, did result in increased task performance (Lerch & Harter, 2001). It seems important to support trainees in understanding feedback that shows performance levels.

Confronting trainees with their error tendencies helps them to avoid making these errors (Dörner & Schaub, 1994). In situations with varying situational dynamics and teamwork interdependencies, the chances to make errors are relatively high. In such situations, errors often appear as a result of a lack of communication or inappropriate task allocation (Sasou & Reason, 1999). To address this, Kontogiannis and Kossiavelou (1999) have made a number of suggestions for making decision support tools more efficient for team decision making. For example, tools should provide information on team-strategy changes fitting to the situation. Furthermore, tools should also provide insight into event escalations and indicate when changes in communication are needed. And finally, they suggest that these tools should indicate when adaptations are needed in the task allocations and structures of team members.

The effectiveness of feedback depends on the timing between task and feedback. First of all, mood, or state-dependent learning shows that retrieval works better when people are in the same mood as they were in when they were learning (Kenealy, 1997). If feedback is delayed, for example after the simulation, it is likely that a person is in a different mood, i.e. no longer stressed. Secondly, trainees will interpret the feedback in the context in which it is given.

Anderson et al. (1995) created a digital tutor and tested four different feedback conditions. In the first condition feedback was given immediately when an error was made and the student was expected to immediately give corrections. In the second condition, feedback was immediate but the student controlled when to provide corrections. In the third condition, the student controlled when feedback was given (not

immediate) and the student controlled the corrections and in the last condition there was no tutor. Students in the first condition showed the most efficient learning. They completed the task the fasted and their results did not differ from the second and third group.

Effective feedback is offered quickly after the task was performed but not during the performance of the skill (Wickens et al., 2004). Trainees will otherwise ignore the feedback resulting in no effect or they will be distracted from the task they are performing which might result in decreased performance. Shute (2008) drew the same conclusion after reviewing the literature on the length and complexity of feedback. When feedback is too long or too complex, trainees will not pay attention to it. Contrary to this finding, there are also studies that did not find an effect of length and complexity of feedback.

This chapter takes the stance that trainees' performances benefit from receiving immediate feedback (from the COPE-FB system described in Chapter 5) about their physiological stress response, predictions about their performance and predictions about the chances of specific errors. These types of feedback are useful in uncertain situations. They let users recognize their current stress state and their behavioural consequences of stress. The COPE-FB system proved effective in the previous chapter, when provided to Naval students working in a high-end simulator. Since providing all three feedback types at once enhanced performances, the study presented here, continues in this line of study by exploring the effects of different combinations of immediate feedback on task performance.

7.1.1. PROTOTYPE EVALUATION

Although the COPE-FB system described in Chapters 5 and 6 is a generic system, different scenarios require different predictive functions. Previous work for example, showed that variables in the COPE model are influenced by tasks characteristics (Chapter 3) and different levels of expertise and prior knowledge can also influence the decision-making process as well as the degree of influence of stress on performances. Therefore, the first experiment of this chapter calibrates the predictive functions. The first experiment also creates a set of stressful scenarios for the second experiment. In the second experiment, the COPE-FB system is used to examine the effect of providing various types of immediate feedback by testing the following three hypotheses:

1. Immediate feedback improves trainees' performances and the perceived usability of the feedback system.
2. Immediate (a) physiological feedback, (b) predicted performance feedback, or (c) predicted error-chance feedback improves trainees' performances and the perceived usability of the feedback system.
3. Offering combinations of immediate feedback types, results in an additional positive contribution on top of the effects created by individual types of feedback, on the trainees' performances and the perceived usability of the feedback system.

EXPERIMENT 1: MODEL PARAMETRIZING

The first experiment was set out to calibrate the predictive models for the specific tasks and target groups in this article. Predictive models were necessary to ran the feedback system during the second experiment. Calibrating the models, in this case, meant determining the parameters in the model. The study was approved by the ethics committees of both TNO Soesterberg and Delft University of Technology.

7.2. METHODS

7.2.1. PARTICIPANTS

Nine participants between 21 and 29 years old, with an average of 24 years old, participated in the experiment. Two of the participants were male. Eight of the participants were interns at TNO and all nine were students at the University of Utrecht. They were all experienced computer users and were naïve with respect to the purpose of the experiment until the debriefing.

7.2.2. EXPERIMENTAL TASK

Since the COPE model was validated with data collected on a Naval ship simulator (Cohen et al., 2015), a task with a similar context was used for this experiment. The task was based on the work of Schreuder and Mioch (2011). During this task, participants saw the layout of a ship on a computer screen as shown in figure 7.1. On the ship, two types of fires would occur: regular fires (1), represented by a white icon, as well as urgent fires (2), represented by a red icon. A normal fire had a timer that indicated how much time there was for extinguishing the fire. In figure 7.1, the user had 70 seconds left to handle the normal fire. Urgent fires did not have a visible timer and burned down faster than normal fires, which meant that urgent fires had to be handled as fast as possible. To bring about extra stress, urgent fires were accompanied by an alarm. When a normal fire had 15 seconds left, participants were warned by a different alarm. The participants needed to collect information about the situation that could be used to determine how to extinguish the fire. This was done by selecting the fire icon followed by selecting a question (3). The answers (yes or no) would appear next to the questions.

A decision tree (figure 7.2) that was handed out to the participants showed the same questions. With the help of the decision tree, participants followed the questions and answers and would end at specific actions (4) that are needed to extinguish the current fire. Decision trees were needed for the participants to know what the right action would be for certain fires. There were four different decision trees. After four scenarios a new decision tree was used to prevent that the task could be executed automatically. The different decision trees had slightly different questions but the structure of the tree would remain.

Figure 7.1 Screenshot of the experimental task. Fires occurring at a ship that needed to be extinguished. Information was gathered to find out how it should be extinguished.

When a fire was selected the consequences of the fire were shown in the form of the number of (virtual) lives at stake (5). This was expected to increase the perceived stress. If a fire was extinguished, all these lives were saved. When a fire was not extinguished, all lives at stake were lost. When the workload became too high, the participant could ask for assistance (7). If the assistance option was selected for a particular fire, it disappeared from the screen. When assistance was used, this option could not be used for the next thirty seconds. This action resulted in loss of lives to prevent participants to ask assistance for all fires. Furthermore, the assistant was not able to handle urgent fires. Some fires also required medical assistance, and the participant needed to notify the sickbay (6).

7.2.3. EXPERIMENTAL SCENARIOS

In this experiment, scenarios were played, which consisted of several fire extinguishing tasks. Scenarios were generated using a scenario generator that was given a set of parameter values. Every scenario had three parameters that could be adjusted to alter the stress induced by the scenario. These parameters were time pressure, information uncertainty and consequences of the decisions. The parameters could be either high or low.

Within every scenario, the two types of fires (regular and urgent) could occur. The consequence parameter has two values (high and low) for both types of fires. High consequences for urgent fires had eight lives at stake and for regular fires six lives were at stake. For low consequences there were two and four lives at stake for respectively

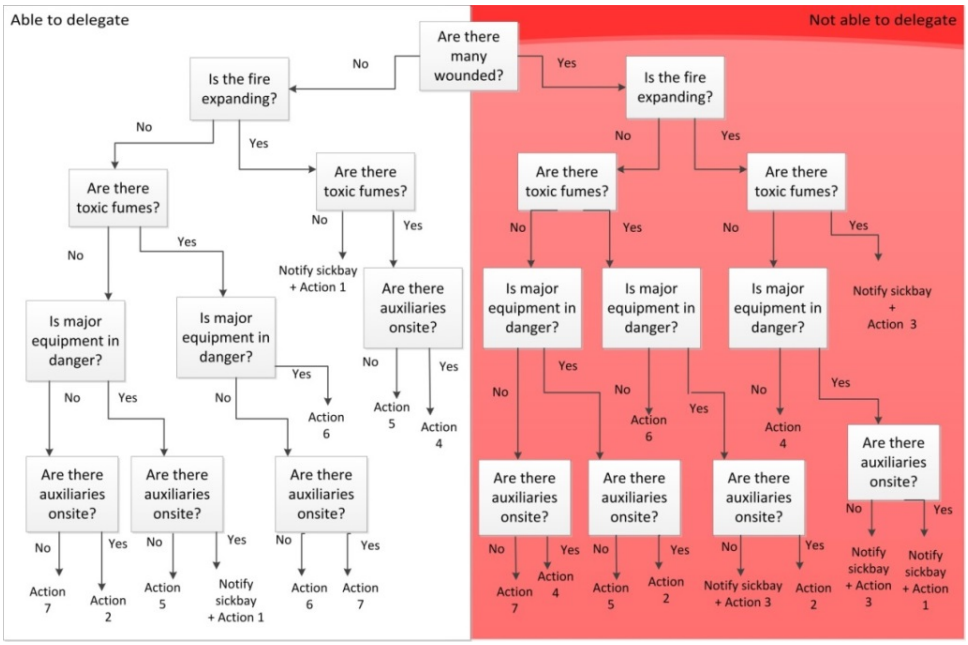


Figure 7.2 Decision-tree for the fire-task. There are 4 decision-trees that were swapped after 4 scenarios. This prevented the participants from performing the task automatically.

regular and urgent fires. For time pressure, there was also a distinction between regular and urgent fires. Time pressure for regular fires was 90+ seconds and for urgent fires the fire needed to be extinguished within 30-50 seconds. High uncertainty meant that answers to a question appeared on average after four seconds whereas low uncertainty gave answers on average after two seconds. For this parameter, no distinction was made between regular and urgent fires.

The combination of all three parameter settings resulted in eight scenarios. In this experiment, participants experienced every parameter setting twice. In other words, they experienced 16 scenarios with 8 parameter settings. The order of scenarios was randomized for each participant. Each scenario lasted for about three minutes each, during which participants had to fight all fires that appeared.

7.2.4. MEASUREMENTS

The COPE variables of appraisal, perceived task demand and arousal were measured for every scenario. Figure 7.2 shows how the work content variables in the COPE model influence the cognitive and affective variables and thereby affect performances. The following subsections explain the variables from figure 7.2 that are measured in this experiment to determine the predictive model parameters.

To determine which scenario was stressful enough to provoke stress in the participants, perceived stress was measured separately together with the Cognitive Task Load measures (Neerincx, 2003).

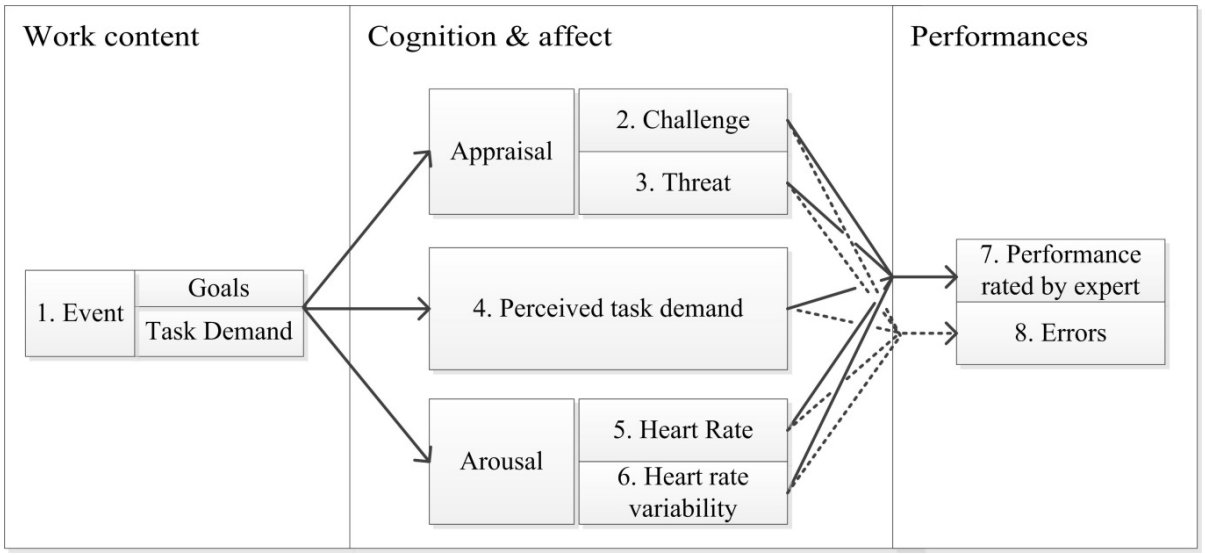


Figure 7.2 COPE model and variables measured in this experiment.

APPRAISAL (2-3)

For measuring the appraisal that was experienced by the participants, a single 10-point scale was used. The question "I experienced the task as…" was answered from "threatening" (1) to "challenging" (10).

PERCEIVED TASK DEMAND (4)

After finishing a scenario, the perceived task demand was measured using the Level-of-Information Processing (LIP) scale from the Cognitive Task Load model (Neerincx, 2003). This model consists of two other levels: Time Occupied (TO) and Task Set Switches (TSS). However, the experimental tasks were constructed in a way that the variables highly correlated, and the TO and TSS measures were incorporated in the experimental tasks (number of fires during one scenario and time pressure). In other words, the diagonal from low to high load was investigated, so that one demand indicator seemed to fulfil. The TO and TSS levels were measured as well, and used to select scenarios as described in Section 7.2.2.

EMOTIONAL STATE: AROUSAL (5-6)

Electrocardiograph (ECG) was recorded with the Zephyr HxM. This is an unobtrusive device attached to a belt, generally used during sport. Participants placed the belt under their clothing on their chest and it sent ECG data via Bluetooth to a laptop. Heart rate (HR) in beats per minute and heart rate variability (HRV) with root-mean-square-successive-differences (RMSSD) were calculated every 10 seconds.

PERFORMANCE (7-8)

Two types of performance were measured: performance score and errors made. The experiment task kept track of the number of lives saved and fires extinguished. Points were rewarded when a fire was extinguished and reduced when a fire was failed to extinguish. The scoring table is shown in table 7.1.

Another measure for performance was the number of errors made during a task. The design of the task allowed four types of errors to occur: communication errors, planning errors, speed errors and task allocation errors. For some fires, participants needed to notify the sickbay. When this action was forgotten or not performed, or when participants did not ask the right number of questions before selecting an action, a communication error occurred and one life was lost. Incorrectly asking for assistance would result in a task-allocation error. When there was an urgent fire but participants handled the regular fire first, a planning error was registered. When participants needed more than 1.25 times the average time to handle fires in a similar situation, a speed-error was registered. The average time to handle fires was calculated after the first experiment. Note that in this experiment participants did not receive immediate feedback about their errors or performance.

PERCEIVED STRESS

Next to the measurements of the COPE model, every scenario was rated by the participants on its stressfulness and difficulty. Perceived stress was measured with one direct question: "How stressful was this scenario". It was answered on a single 5-point scale ranging from not stressful, to very stressful.

Per task, the participants filled in the Cognitive Task Load (CTL) questionnaire from Neerincx (2003). The 'Level of Information Procession', 'Time Occupied' and the 'Task Set Switches' were rated on a 5-point scale for every task.

Table 7.1. Performance scoring scheme for different actions.

Action	Low consequence fire	High consequence fire
Asking for help correctly	-1	-1
Asking for help incorrectly	- all lives at stake	- all lives at stake
Notify sickbay when needed	0	0
Forget to notify sickbay	lives saves or lost -1	lives saved or lost -1
Extinguish a regular fire	+ 2	+ 2
Extinguish an urgent fire	+ 4	+ 6
Burn down a regular fire	-2	-3
Burn down an urgent fire	-4	-6

7.2.5. *PROCEDURE*

At arrival, the participants were asked to put on the heart rate monitor. The participants then read the experimental and the task instructions while the experimenters checked if the heart rate monitor was working. Questions about the instructions could be asked before a tutorial trial was started with a printed version of the first decision tree. This tutorial showed the participants how to perform the task. When the task was understood, the experimental trials started. After every single scenario, participants filled in the questionnaires for the appraisal and task demand. After every four scenarios the decision tree was changed for another tree that had slightly different questions. Multiple decision trees were created to prevent participants from automatically selecting the order of the questions without reading them. A questionnaire with demographic information was filled in at the end of the experiment. This experiment lasted between 90 to 120 minutes.

7.3. RESULTS

Data from this experiment was used to select the most stressful scenarios. These scenarios were used in the second experiment. The predictive functions were created based upon COPE variable data from these scenarios.

7.3.1. *DATA PREPARATION*

For every scenario, heart rate data, heart rate variability data and the questionnaire data were collected. Arousal data was collected every 10 seconds. Since the scenarios lasted approximately 3 minutes, this led to a list of about 18 data points per scenario. The appraisal and task demand values per scenario were the average values given by the participants in experiment 1.

7.3.2. *SCENARIO SELECTION*

Every scenario was rated on perceived stress and the measures of CTL. The three CTL levels and the perceived stress score were used to determine the overall stressfulness of the scenarios. The median score was calculated for these variables for every scenario. Data of the scenarios with the same parameters were then combined by averaging the median scores. These average scores were used to rank the scenarios on 4 levels. Scenario 6 (*) had the highest scores on CTL, but did not score high on perceived stress. Participants had saved fewer lives in scenario 6 than in other scenarios which led to believe that scenario 6 might have been too difficult and participants gave up, which can explain feeling less stress and the decrease in performance. Therefore, scenario 6 was

not selected for experiment 2. Scenarios that were selected were 8, 2, 4 and 5 since they scored high on perceived stress and on the CTL levels. These scenarios were attached to their equivalent scenario (16, 10, 12 and 15) to create new scenarios for experiment 2. Every pair created 2 new scenarios for the second experiment. For example; scenarios 8+16 and scenarios 16+8 were two new scenarios.

7.3.3. PREDICTIVE MODELS

Due to technical problems, performance data of one participant was lost which meant that none of the data of this participant could be used for the calibration. Thus the COPE-FB System was calibrated using data of 8 participants of which 2 were male. Four Generalized Linear Mixed Models (GLMM) were created in SPSS 20.0. One model predicted performance using a linear model and three models predicted different errors using a binary logistic regression model. No planning and errors concerning the speed of the task execution were found in the dataset. Specific models could therefore, only be made for communication and task allocation errors. The fixed factors consisted of the COPE variables: HR, HRV, challenge, threat and perceived task demand. A participant-factor was included as a random factor to control for participant variation. The random effect covariance type was set to Variance Component.

PREDICTIVE PERFORMANCE MODEL

A GLMM shows that the fixed factors could explain the performance, ($F(5,24.44) = 24.23$, $p < 0.05$) with a weak Spearman rho correlation of $r = 0.12$ between observed and predicted performance. The individual variance did not differ from the standard intercept (varintercept = 0.092, Std Err = 0.109, Z = 0.844, p. = 0.399), indicating that on average the participants did not differ in their performance. The coefficients in table 7.2 show that a decrease in heart rate and an increase in heart rate variability coincided with an increase in standardized performance. Additionally, an increase in challenge and perceived task demand coincided with an increase in the standardized performance.

PREDICTIVE ERROR MODELS

Before predictive models could be made for the error variables, the underrepresentation of errors compared to no-errors in the dataset needed to be corrected. The error variables are binomial (0 = no error and 1 = error) and the observed ratio of all errors was skewed towards 0. The total error ratio was 888:30 (29.6:1), for communication errors this was 904:14 (64.57:1) and for task allocation it was 902:16 (56.38:1). By weighing the data, the ratios were stretched towards a 10:1 ratio. This ratio was chosen since it still showed a favour for 'no errors'. The total error cases were weighted with the ratio of 25:75. For the communication and task allocation errors a ratio of 15:85 was used. After applying these weightings, the new ratios were 9.87:1, 11.39:1 and 9.95:1 for

Table 7.2. Predictive performance model. The model consists of 5 variables.

Variables	Coefficient	Std. Error	*t*	*p*
Intercept	-0.028	0.081	-0.35	0.727
Arousal: HR	-0.003	0.001	-2.86	0.004
Arousal: HRV	0.013	0.003	3.95	<0.001
Appraisal: challenge	0.033	0.009	3.54	<0.001
Appraisal: threat	0.009	0.006	1.57	0.116
Perceived Task Demand	0.048	0.008	6.36	<0.001

the total errors, communication errors and task allocations, respectively. The predictive models were based on the weighted dataset.

The predictive model for the total error category is shown in table 7.3 and is able to predict errors based on HRV, challenge and level-of-information processing $F(5, 24.44) = 17.46$, $p<0.05$. An ROC-curve for this model provided an Area Under the Curve (AUC) value of 0.725 which states the model is reasonably well in discriminating between errors and no-errors (Swets, 1988). The individual variance did not differ from the standard intercept (varintercept = 0.451, Std Err = 0.526, Z = 0.858, p. = 0.391), indicating that on average the participants did not differ in their performance.

Predictions for communication and task allocation errors can be made out of the models shown in table 7.4. Communication errors could be predicted out of all the variables ($F (5, 14.744) = 52.566$, $p<0.05$). An ROC-curve provided the Area Under the Curve (AUC) value of 0.790 for this model. The individual variance did not differ from the standard intercept (varintercept = 8.80, Std Err = 11.555, Z = 0.761, p. = 0.447), indicating that on average the participants did not differ in their performance.

Task allocation errors could be predicted out of HRV, challenge, threat and level-of-information processing ($F(5, 14.884) = 51.78$, $p<0.05$). An ROC-curve provided the Area Under the Curve (AUC) value of 0.665 for this model. The individual variance again did not differ from the standard intercept (varintercept = 2.474, Std Err = 3.367, Z = 0.735, p. = 0.463), indicating that on average the participants did not differ in their performance.

Table 7.3. Logistic regression model to prediction the total errors. Errors are weighted with 25:75.

	Coefficient	Std. Error	*t*	*p*
Intercept	3.233	0.461	7.02	<0.001
Arousal: HR	-0.005	0.004	-1.27	0.204
Arousal: HRV	-0.045	0.009	-5.18	<0.001
Appraisal: challenge	0.215	0.034	6.27	<0.001
Appraisal: threat	0.016	0.019	0.88	0.378
Perceived Task Demand	-0.191	0.029	-6.70	<0.001

Table 7.4. Logistic regression model to prediction communication and task allocation errors[1].

Error type	Coefficient	Std. Error	*t*	*p*
Communication Errors				
Intercept	9.384	1.783	5.26	<0.001
Arousal: HR	-0.026	0.009	-3.01	0.003
Arousal: HRV	-0.093	0.011	-8.82	<0.001
Appraisal: challenge	-0.451	0.057	-7.90	<0.001
Appraisal: threat	-0.192	0.031	-6.11	<0.001
Perceived Task Demand	-0.389	0.053	-7.35	<0.001
Task allocation Errors				
Intercept	1.626	0.912	1.78	0.075
Arousal: HR	0.000	0.004	0.08	0.938
Arousal: HRV	0.725	0.269	2.70	0.007
Appraisal: challenge	0.609	0.043	14.04	<0.001
Appraisal: threat	0.124	0.023	5.45	<0.001
Perceived Task Demand	-0.098	0.035	-2.77	0.006

[1]These errors are weighted with 15:85

7.4. DISCUSSION

The first experiment resulted in a set of 8 stressful scenarios that will be used in the next experiment. Based on data from these scenarios, four significant models were created that predicted performance, communication errors, task allocation errors and total number of errors out of the variables of the COPE-model. The predictive models for performance and for total number of errors are shown in figure 7.3. Challenge (2), perceived task demand (4), HR (5) and HRV (6) were predictors for performance (7) (solid lines) whereas challenge (2), perceived task demand (4) and HRV (6) were predictors for the total number of errors (8) (dotted lines).

Some aspects of the parameter settings however are open for discussion. For example, the error models were based on a weighted dataset. The weightings changed the error ratios to 1:10 error, no-error ratio. Whether the ratios chosen in this experiment were satisfying depends on the interpretation of the consequences of misses and false positives in the error prediction. Since the feedback system will be used in training settings the consequences of false positives and missing errors are not severe.

Another discussion point is the 10 seconds interval in which the physiology was measured. This window was chosen because the feedback would be used in tasks that last approximately 3 minutes. If feedback changed every minute, only three moments of feedback will occur. Therefore, it seemed appropriate to set the predictions, and therefore the dataset, to 10 seconds to increase the amount of feedback moments.

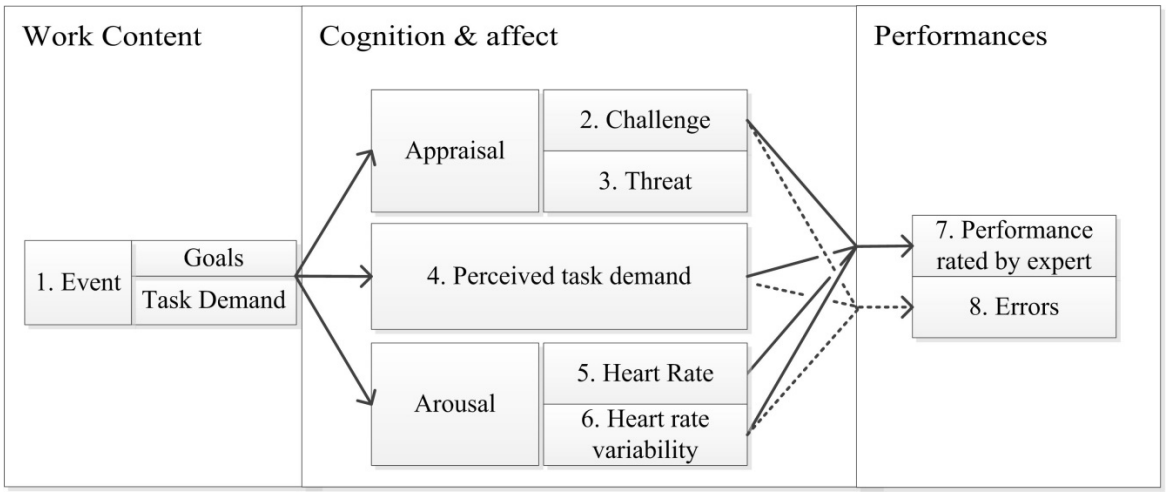


Figure 7.3 Graphical presentation of the COPE model and how performance and error chances can be predicted out of these variable.

For the next experiment, these models were implemented into the COPE feedback system. This system used input variables and the predictive models to calculate performance and chances of specific errors that were shown to the user this time.

EXPERIMENT 2: FEEDBACK TEST

The second experiment focussed on analysing the impact that different feedback types have on performance and their perceived usability, when trainees work in a stressful virtual training setting. The experiment was setup as a within-subjects design. Participants were provided with or without (1) physiological feedback, (2) performance prediction feedback, or (3) error-chance prediction feedback. Using a full-factorial design (2×2×2), participants were exposed to eight different feedback conditions. Again, the experiment was approved by the ethics committee of TNO Soesterberg and Delft University of Technology.

7.5. METHODS

7.5.1. PARTICIPANTS

A total of 29 participants were recruited from a participant database at TNO, a research institute in the Netherlands. People who had participated in the previous experiment, were excluded. Participants were between 18 and 34 years old, with an average of 25.5 (SD = 4.67) years. Fifteen participants were male and all participants were naive with respect to the purpose of the experiment. They were compensated with 25 euros plus travel expenses. A bonus of 20 euros was awarded to the participant with the highest performance score on the experimental task.

7.5.2 TASK

The same fire extinguish task was used as in the first experiment. The participants were confronted with the eight scenarios selected in the first experiment. While the scenarios were being performed, the participants received eight different immediate feedback combinations via the COPE-feedback system.

7.5.3 USING THE COPE-FB SYSTEM

The models used by the COPE-FB system need five real-time input values to calculate the performance predictions and error-chance predictions. Heart rate and heart rate variability were measured real-time with the Zephyr HxM. Appraisal (challenge or threat) and task demand were rated per task in experiment 1. The experimenter selected the scenario that was performed, and the COPE-FB system read a file that contains the values rated in experiment 1. The models output was used as the feedback. The feedback was updated every 10 second.

The COPE-FB system consisted of two parts, the trainer part and the trainee part. The trainer part was used by the trainer, or in this case the experimenter. Figure 7.4 shows a screen shot of the trainer part. First, files were selected containing regression models (1) and scenario variable values (2). In the third step, the heart rate device was connected via Bluetooth (3) so the system could use the input signal. Next, the trainer

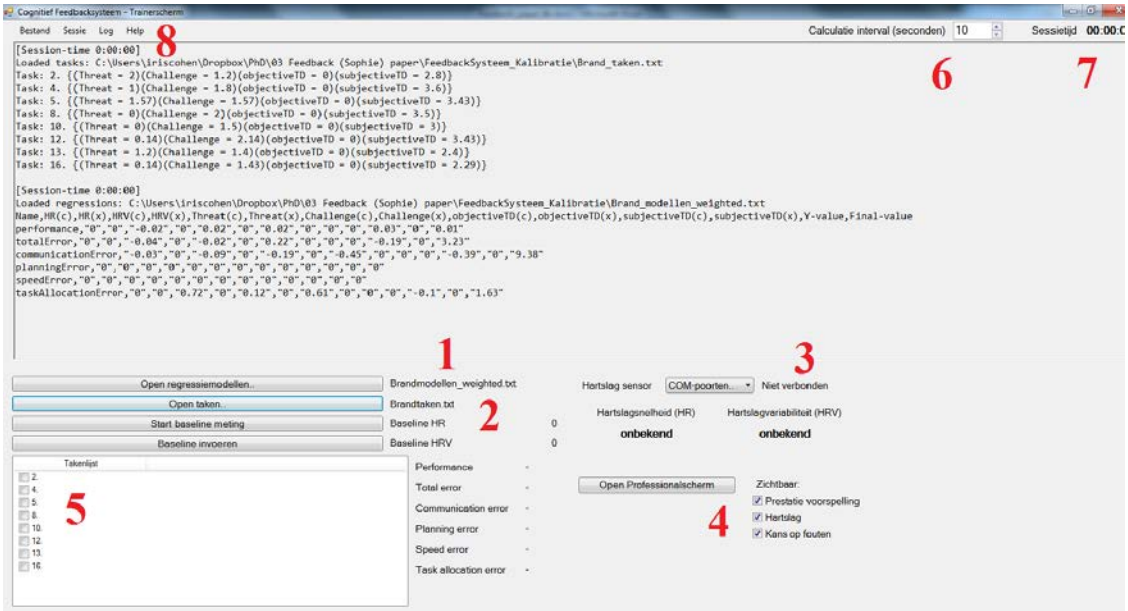


Figure 7.4 The screen for the trainers (a)

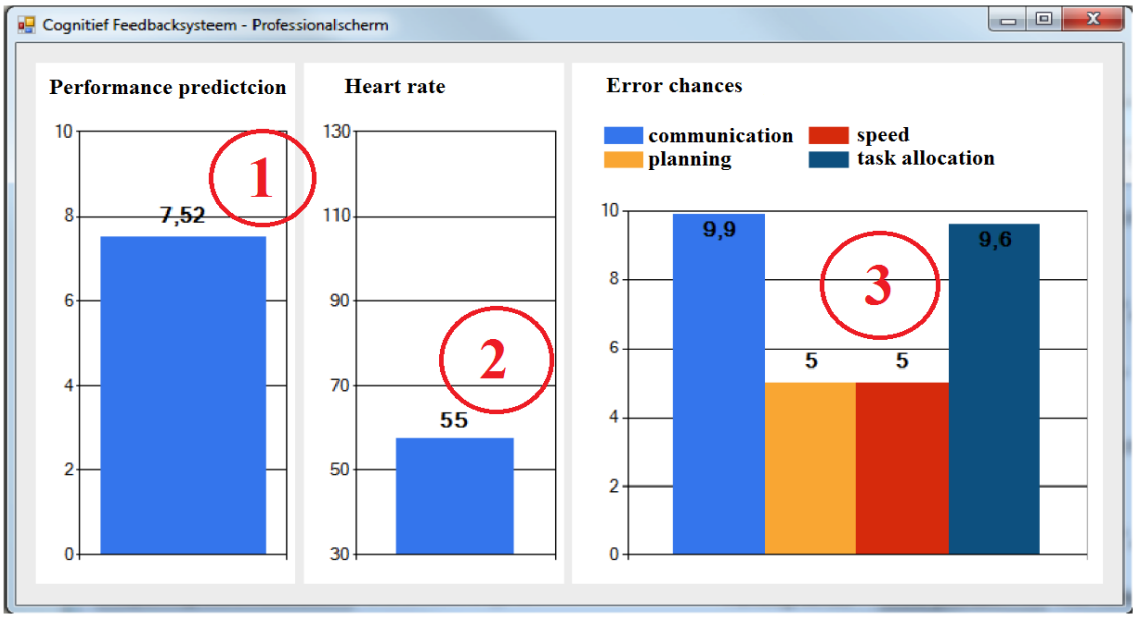


Figure 7.5 Trainee screen from the COPE-FB system showing three types of feedback.

selected which parts of the feedback would be shown to the trainee, i.e. participant (4). When the system was running, the trainer would select which scenario was performed (5). By selecting a scenario, the appraisal and task demand values for that scenario were sent from the scenario file to the regression models. Additional functions that were available in the trainer-component were: setting the calculation time-interval (6) for new predictions and feedback; tracking the current session time (7); and an option for immediate logging of the data (8).

The trainee screen only showed the output of the predictive models divided over three panels with bar graphs as shown in figure 7.5. On the left, the performance prediction was shown in one bar graph. In the middle, a bar graph showed the trainees current heart rate. On the right, the predicted error chances were shown. By default, four bar graphs were shown. Above these graphs, the legend showed which graph colour corresponds to which error.

The first experiment in this chapter explained that there was no data for planning and speed errors and therefore, no predictions about these errors could be made. The bar graphs in the feedback screen for those two errors remained therefore static on 5 (as shown in figure 7.5). If these graphs would be set to 0, participants might have thought that they were not making these errors. The participants were told that these errors would not be predicted.

7.5.4 MEASUREMENTS

PERFORMANCE AND ERRORS

There were two measures for performance: the total score for a task, and the number of errors made. The scoring table and error categorization were as in the first experiment.

USABILITY

After every scenario, participants were asked to judge the usability of the feedback screen provided by the COPE-FB system during that specific scenario by filling in the System Usability Scale (SUS). The SUS consisted of 10 items about the systems usability that were answered on a 5-point scale (Brooke, 1996). To calculate the total SUS score, items 1, 3, 5, 7 and 9 were scored by subtracting 1 from the scale position. For items 2, 4, 6, 8 and 10, the scale position was subtracted from 5. These scores were then summed and multiplied by 2.5 to obtain an overall value for SUS. After transposing, the SUS scores have a range from 0 to 100.

Next to the SUS, participants were asked to choose one of the feedback types as the most pleasant and one as the least pleasant type of feedback. They were also asked to indicate why they chose those types in an open question format.

OTHER MEASURES

Figure 7.6 shows a participant in the experimental setting. Note that she is wearing more sensors than just the Zephyr HxM heart rate belt. For another experiment, facial movement was measured using electromyography (EMG). Data from these sensors did not enter the analyses of this chapter. These sensors influenced all the participants the same manner throughout the eight conditions.

7.5.5 PROCEDURE

At arrival, the participants put on the Zephyr HxM, read instructions and signed a consent form. The instructions consisted of an explanation of the task and the feedback system. A tutorial was started to practice the task and learn about all the options of the ship simulator. Next, the feedback screen was turned on simultaneously with the data-recording session of the COPE-FB system, and finally the first scenario was started. The order of the eight different feedback conditions was counterbalanced as shown in table 7.5. The scenarios were all executed by the participants in the same order. After finishing each scenario, participants filled in the questionnaire about the scenario and the COPE-feedback system. After every four scenarios, the decision-tree was exchanged for another decision tree. A total of eight scenarios were performed. After the experimental task, demographic information was collected and the participants chose their most favourite and least favourite type or combination of feedback.

Figure 7.6 Experimental setup. In the foreground trainer part of COPE-FB system with Zephyr HxM heart on keyboard. In the background a participant views the ship simulation on the left and the trainee part of COPE-FB system on the right.

Table 7.5. Orders of experimental conditions. The scenario order did not change during the experiment.

Pf:Er	Control	ER	HR:Pf	HR	HR:Er	Pf	HR:Pf:Er
Control	Pf	HR:Pf	HR	HR:Er	Pf:Er	HR:Pf:Er	Er
HR:Pf	Pf:Er	HR	Control	HR:Pf:Er	Pf	Er	HR:Er
HR:Er	HR:Pf:Er	Pf	Pf:Er	Control	Er	HR	HR:Pf
Er	HR:Er	Pf:Er	HR:Pf:Er	Pf	HR:Pf	Control	HR
Pf	HR:Pf	HR:Pf:Er	Er	Pf:Er	HR	HR:Er	Control
HR	Er	Control	HR:Er	HR:Pf	HR:Pf:Er	Pf:Er	Pf
HR:Pf:Er	HR	HR:Er	Pf	Er	Control	HR:Pf	Pf:Er

HR=heart rate feedback, Pf=performance prediction feedback, Er=error-chance prediction feedback

7.5.6 DATA PREPARATION AND ANALYSES

With SPSS 20.0, heart rate outliers were detected and removed from the data file. Heart rate data was considered an outlier when it was deviating more than 2.5 times SDs from the mean. One participant had more than 25% of the heart rate data discarded and this person was therefore excluded from all analyses. The feedback, based on physiological values, for this person was unlikely to have been optimal. The usability analysis was therefore also based on a sample of 28 participants, of which 14 were male.

For the performance-analysis, the data of another three participants was discarded. During their experimental sessions technical issues with the COPE-FB system resulted in

incorrect performance and error scores. This analysis was therefore based on a sample of 25 participants, of which 14 were male.

A relative performance score was used in the analyses. This score was calculated by dividing the 'lives saved in a condition', by the 'total number of lives that could have been saved in that condition'. The descriptive statistics for the performance scores and the relative performances scores are shown in table 7.6. There were two types of specific errors measured during the tasks: communication and task allocation errors. Adding these two resulted in total number of error. Figure 7.7 shows that the distributions of the three error variables resemble a Poisson distribution.

Table 7.6. Descriptive statistics for the (relative) performance scores.

	Performance scores	Relative performance scores
Minimum	-30	-0.94
Median	8.50	0.27
Mean	6.64	0.21
Maximum	32	1

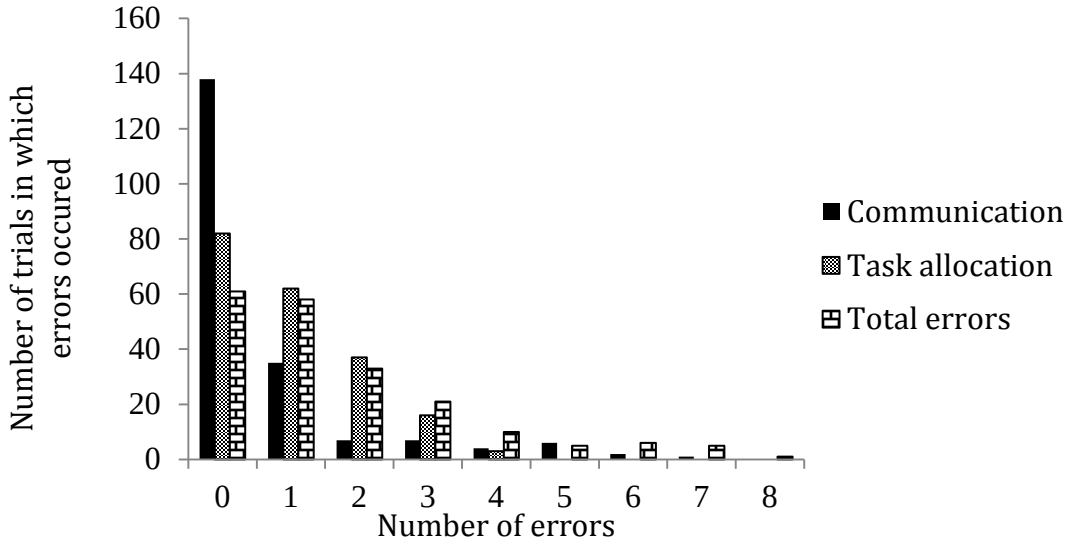


Figure 7.7 The number of trails in which zero to eight communication, task allocation and total errors were observed. Total errors show errors without dividing them over specific categories (either communication or task allocation). These bar graphs show that the error variables are not normally distributed.

The statistical analyses were executed in R studio. The effect for the different feedback conditions was examined using a linear mixed-effect model (LMER) function on the performance and SUS data, and a generalized linear mixed-effect model (GLMER) with the Poisson family function for error data. LMER fits linear-mixed effect models to data whereas GLMER fits generalized linear mixed-effect models to datasets. The ordinal preference data was analysed using an exact multinomial and exact binomial tests from the EMT package in R.

7.6 RESULTS

The analyses focus on the improvements of performance, number of errors, and perceived usability measured for all the feedback conditions. The first hypothesis states that immediate feedback in general results in an increase of performance and perceived level of usability. The second hypothesis states that the three separate feedback types increase performance and perceived level of usability. The third hypothesis states that an additional positive effect can be found on top of the effect for the separate feedback types on performance and perceived level of usability. The presentation of the results follows the order of the hypotheses. First, all feedback conditions were examined as a two level factor: no feedback and feedback. Next, the feedback conditions were examined separately.

7.6.2 PERFORMANCE

Two models were created with performance as a dependent variable: a null model with no fixed factors, including only a random intercept factor for participants, and an alternative model that added a fixed two level factor (feedback, no-feedback) to the null model. A likelihood ratio test found that the model fit of the alternative-model was an improvement over the model fit of the null model ($\chi 2(1) = 5.38$, $p = 0.02$).

Table 7.7. Likelihood ratio test for models fitting the performance scores. Testing if adding factors will improve the fit compared to the H0 model.

Model	*df*	Log likelihood ratio	χ^2	*df*	*p*
1. H0 model	3	-82.72			
2. 1 + main effects	6	-82.01	1.43	3	0.699
3. 2 + 2way interactions	9	-81.03	3.39	6	0.759
4. 3 + 3way interaction	10	-79.23	6.99	7	0.431

Therefore, the alternative model explains more than just an average performance score. Relative performance scores when no feedback was provided (M = 0.07, SD = 0.49) was lower than the relative performance scores when feedback (M = 0.23, SD = 0.40) was provided (t(198) = 2.32, p = 0.021).

Next, the feedback factor was split into the different feedback types. Three extra models were created that contained either (2) only the main effects of heart rate, performance predictions and error chance predictions, (3) the main effects and the 2-way interactions, and (4) the main effects, 2-way and 3-way interactions for three types of feedback. As table 7.7 shows, adding the three main factors and interaction factors did not improve the model fit compared to the null model. In other words, no significant effect was found for the main effects or the interaction effects.

7.6.3 ERRORS

The first step of the error analysis was again to test whether feedback in general resulted in any error reduction. Again a null model and an alternative model with fixed two-level factor (feedback, no-feedback) were created. The fit of the null model did not improve when a feedback factor was added for the communication errors $\chi2(1) = 2.62$, $p = 0.11$, the task allocation error $\chi2(1) = 0.11$, $p = 0.74$, or total number of errors $\chi2(1) = 1.59$, $p = 0.21$.

Table 7.8. Likelihood ratio test for models fitting the Communication, task allocation and total error variable. Testing if adding effects would improve fit compare to the H0 model.

Error Type	*df*	Log likelihood ratio	χ^2	*df*	*p*
Communication error					
1. H0 model	2	-208.33			
2. 1 + main effects	5	-205.79	5.0746	3	0.1664
3. 2 + 2way interactions	8	-204.69	2.2051	6	0.2958
4. 3 + 3way interaction	9	-202.63	4.1163	7	0.1223
Task allocation error					
1. H0 model	2	-243.04			
2. 1 + main effects	5	-242.86	0.3470	3	0.9510
3. 2 + 2way interactions	8	-241.38	3.3245	6	0.7672
4. 3 + 3way interaction	9	-241.27	3.5368	7	0.8313
Total error					
1. H0 model	2	-208.33			
2. 1 + main effects	5	-205.79	5.0746	3	0.1664
3. 2 + 2way interactions	8	-204.69	7.2797	6	0.2958
4. 3 + 3way interaction	9	-202.63	11.396	7	0.1223

. p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Next, the feedback factor was again split into the separate feedback types and combinations. Additional models were created that contained either (2) only the main effects of heart rate, performance and error chance predictions, (3) the main effects and the 2-way interactions, and (4) the main effects, 2-way and 3-way interactions for three types of feedback. Again, these models did not improve model fit compared to the null model as shows in table 7.8.

7.6.4 USABILITY

The usability of the different feedback conditions of the COPE-feedback system were measured with the System Usability Scale and with a rating scale on which feedback was most pleasant and which one was least pleasant.

SUS SCORES

As before, the first step was to analyse the effect of feedback in general. Two models were again created to fit the SUS scores, a null-model and an alternative model including feedback as two-level factor. A likelihood ratio analysis found that the model fit of the alternative model was no improvement over the model fit of the null model ($\chi 2(1) = 2.66$, $p = 0.10$).

The second analysis step examined whether individual types of feedback or their interactions affected SUS scores. Four models were created as was done with analysis of performance and error data. All models significantly improved the fit compare to the null model (table 7.9).

The fourth model, including main effects, 2-way interaction effects and 3-way interaction effects is analysed and presented in table 7.10. The SUS score without HR feedback (M=51, SD=14) was significantly lower ($t(111) = -10.77$, $p<0.05$) then the SUS score when HR feedback was provided (M=55, SD=16). This main effect is illustrated in figure 7.8a.

Table 7.9. Likelihood ratio test for models fitting the SUS variable. Testing if adding effects will improve the H0 model.

Model	*df*	Log likelihood ratio	χ^2	*df*	*p*
1. H0 model	3	-897.69			
2. 1 + main effects	6	-892.90	9.58	3	* 0.022
3. 2 + 2way interactions	9	-886.80	21.77	6	** 0.002
4. 3 + 3way interaction	10	-886.28	22.82	7	** 0.002

. $p<0.1$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Table 7.10. Effects of feedback types on SUS scores; Model 4 including the main effects, 2-way interactions and 3-way interaction.

Effects	*df*	Sum of Squares	*F*	*p*
HR	1	1017.89	7.75	** 0.006
Performance	1	4.72	0.04	0.850
Error	1	307.62	2.34	0.128
HR × Performance	1	26.81	0.20	0.652
HR × Error	1	1265.88	9.64	** 0.002
Performance × Error	1	307.62	2.34	0.128
HR×Performance×Error	1	132.84	1.01	0.316

. p<0.1, * p<0.05, ** p<0.01, *** p<0.001

T-tests investigating the interaction effect (shown in figure 7.8b) showed as could be expected from the main effect that the SUS score for no feedback (M=50, SD=10) was significantly lower (t(27) = -3.799, p=0.001) than for the only heart rate feedback (M=59, SD=13) condition. Likewise, the SUS scores for the only error feedback (M=52, SD=16) condition were also significantly lower t(27)=-2.388, p=0.024 than for the only heart rate feedback condition. However, the same significant difference (t(27)=2.676 , p=0.013) was found between the only heart rate feedback condition and the heart rate combined with error feedback condition (M=52 , SD=14). This suggests that adding error feedback to heart rate feedback lowered again the perceived usability.

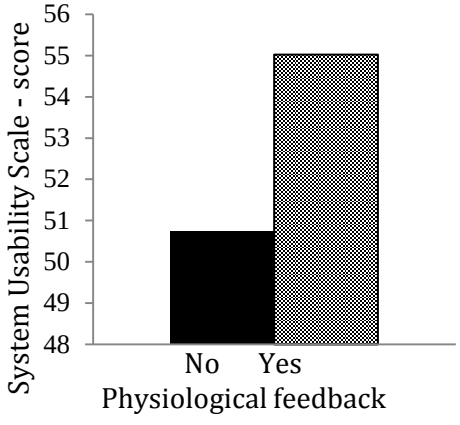


Figure 7.8a) The main effect of HR feedback on SUS score.

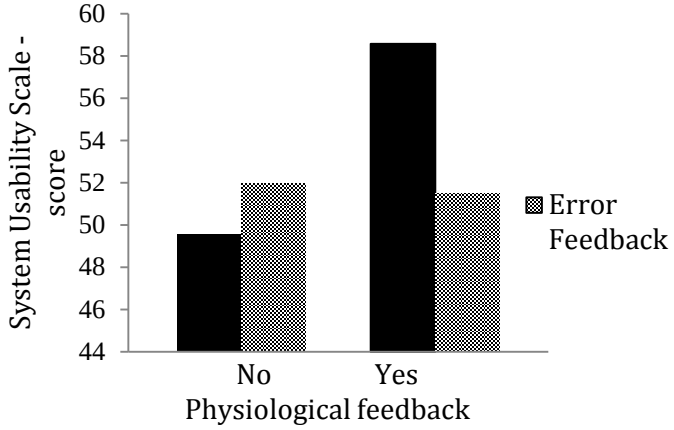


Figure 7.8b) The interaction effect of HR×Error feedback on SUS score.

PREFERENCES RATING

The participants rated which feedback type they thought was most pleasant and which the least pleasant, and gave a reasoning behind their choice. Figure 7.9 shows in bar graphs how often a feedback condition was chosen as most or least pleasant. With exact multinomial and exact binomial tests, it was tested if the ratings were distributed fairly over all conditions, or if they showed a preference. If there was no preference towards a specific feedback condition, the chances should be equal to 1/8.

The observed ratings showed a significant preference for one of the feedback types for both the most pleasant rating (n=28, expected probability = 0.125, $p < 0.001$) and least pleasant rating (n=28, expected probability = 0.125, $p = 0.003$). The condition without feedback was rated most pleasant by 41.38% of the participants. The other 58.62% choose a type of feedback as most pleasant. An exact binomial test shows that the distribution of most pleasant ratings of feedback versus no-feedback, does not show a preference (n=28, expected probability 0.125:0.875, p=0.345).

For the least pleasant ratings, both the 'no feedback' and 'all feedback' conditions scored the highest score of 25.57%. An exact binomial test with these two options opposed the other six options was conducted. This distribution differs from a random probability distribution (n=28, expected probability 0.125:0.125:0.75, $p<0.001$).

No participant rated the conditions with heart rate feedback combined with performance feedback (n=28, expected probability 0.125:0.875, $p<0.001$) or heart rate feedback combined with error feedback (n=28, expected probability 0.125:0.875, $p<0.001$) as their least preferred option. Related, the feedback condition combining performance prediction and error chance prediction was rated as most pleasant by no one (n=28, expected probability 0.125:0.875, $p<0.001$).

The participants' reasons underlying their preferences were mainly practical ones and are collected in table 7.11. The feedback was distracting them (n=6) or they did not have time to watch the feedback screen (n=3). Explanations concerning the applicability of the feedback stated that participants did not understand the feedback (n=2) or they thought they received too much information (n=2). Surprisingly, reasons for the participants to report feedback as pleasant contradicted the reasons to dislike feedback.

Table 7.11. Participants' reasons to rate a feedback combination as either most or least pleasant.

Least pleasant	*n*	Most pleasant	*n*
No time to watch the screen	3	Useful information	5
Too much distraction	6	I understand this	2
I don't understand it	2	This is useful	2
Too much information	2	I know what to do with it	5
No added value	1	I changed strategy with this feedback	4

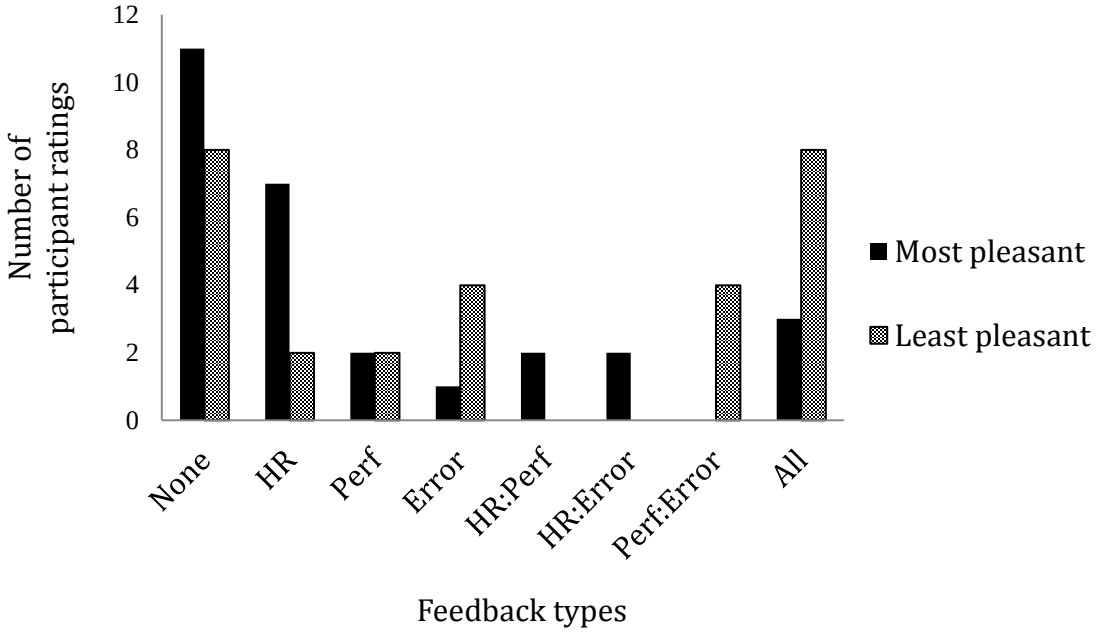


Figure 7.9 Bar graphs show how often participants selected a specific feedback type or combination of feedback, as most pleasant and as least pleasant. HR = heart rate/physiological feedback, Perf = predicted performance feedback, Error = predicted error-chance feedback

Participants rated feedback as useful (n=5), and participants reported that they knew what to do when receiving a certain type of feedback (n=5). Two participants stated that they understood the feedback and four participants even explained that they would change strategy when certain feedback was given.

7.7 DISCUSSION

Support for the first hypothesis was only found in the performance scores. The results showed that the performance scores increased when feedback was presented to the participants. However, no support was found for this hypothesis concerning the number of errors or the SUS scores. Support for the second hypothesis was only found in the analysis of the SUS scores. The SUS score for physiological feedback was higher compared to conditions were physiological was not provided.

Instead of support for the third hypothesis that a combination of feedback makes a positive contribution, the findings provided grounds to, at least partly, reject this hypothesis with regard to perceived usability. Adding error-chance feedback to physiological feedback reduced the perceived usability when comparing it with a situation where only physiological feedback was provided. However, when it came to the effect on performance the findings were inconclusive on this point.

There are some limitations that should be considered when the results of this study are interpreted. One limitation concerns the explanation of the COPE feedback

system to the participants. A digital tutorial was created to explain the experimental task, but for the COPE feedback system, a written explanation was given to the participants. A more in-depth tutorial about the COPE feedback system might help the participants to understand the system better. Furthermore a training session with the COPE-FB system might help participants to get more familiar with the system. Another benefit from a digital tutorial is that the participants will not be able to skip the tutorial.

Another limitation in the experimental design was that the COPE-FB system was designed to predict four types of errors (communication, planning, speed and task allocation errors). These errors originated from a previous study where a more naturalistic task was performed in a more complex virtual environment (Cohen et al., 2015). The computer task in this study was derived from a task that did not naturally evoke these errors. We did enrich this task in order for these errors to occur, but still, participants in the first experiment only made two types of errors. This meant that only two of the four errors could be predicted which meant an incomplete use of the feedback system.

7.8 General discussion

This chapter described the evaluation of the newly created COPE FeedBack system (COPE-FB system), designed to decrease negative effects of stress on performances. The first experiment successfully created stressful tasks and established parameters for predictive models. For the second experiment, the predictive models were implemented into the feedback system to provide participants with eight different combinations of three types of feedback. The statistical analyses showed that providing participants with immediate feedback resulted in an improvement of performance scores. Analysing the main and interaction effects of the different types of feedback showed an increase of SUS score for physiological feedback over no physiological feedback. But it also showed that this improvement could be undone by adding error chance feedback to the error-physiological feedback. The usability data shows that there are good opportunities for this type of feedback to be accepted and processed for performance enhancement. To establish such enhancement, the feedback needs substantial improvements. Participants were interested in seeing their heart rate but seemed confused as to how the other types of feedback worked. A more in-depth tutorial session should be added to the COPE-FB system to increase the understanding of the provided feedback. This will also increase trust in the system which is necessary in order for it to work effectively (Grootjen, Bierman, & Neerincx, 2006).

Another reason for participants not to prefer different types of feedback is that it is too distracting. This problem might be solved with a change in the design. The current version of the COPE-FB system shows consistency in the design of the different types of feedback (Horsky et al., 2012), to rule out design effects between the feedback types. Different designs might be easier and faster to interpret. Another option is

implementing an auditory warning signal when performance decreases to a certain threshold. If users only have to watch the feedback when a signal is provided, the users do not have to decide for themselves when to watch the feedback.

The results of this study are promising. Performance was improved when feedback was provided. Also the preference ratings illustrated that psychological feedback was preferred by the users. Although feedback was often seen as a distraction, users also assessed the information provided as useful.

ACKNOWLEDGEMENT

The work presented in this chapter is supported by the Dutch FES program: Brain and Cognition: Societal Innovation (project no. 056-22-010). Thanks to Maarten van der Smagt for his input as a student supervisor from University Utrecht.

REFERENCES

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. The journal of the learning sciences, 4(2), 167-207.

Andresen, L., Boud, D., & Cohen, R. (2001). Experience-Based learning. In G. Foley (Ed.), Understanding adult education and training (pp. 225-239). Sydney: Allen & Unwin.

Beach, L. R., & Lipshitz, R. (1993). Why classical decision theory is an inappropriate standard for evaluating and aiding most human decision making. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), Decision making in action: Models and methods (pp. 3-20). Norwood, New Jersey: Ablex publishing corporation.

Bouchard, S., Bernier, F., Boivin, E., Morin, B., & Robillard, G. (2012). Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. Plos one, 7(4).

Brooke, J. (1996). SUS: a quick and dirty usability scale. In P. W. Jordan, B. Thomas, I. L. McClelland & B. Weerdmeester (Eds.), Usability evaluation in industry: CRC Press.

Busscher, B., Vlieger, D. d., Ling, Y., & Brinkman, W.P. (2011). Physiological measures and selfreport to evaluate neutral virtual reality worlds. Journal of Cybertherapy & Rehabilitation, 4(1), 15-25.

Cesta, A., Cortellessa, G., & Benedictis, R. D. (2014). Training for crisis decision making - An approach based on plan adaptation. Knowledgde-based systems, 58, 98-112.

Cohen, I., Brinkman, W. P., & Neerincx, M. A. (2015). Modelling environmental and cognitive factors to predict performance in a stressful training scenario on a naval ship simulator. Cognition, Technology & Work, 2015.

Cohen, M. S. (1993). The bottom line: naturalistic decision aiding. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), Decision making in action: Models and methods (pp. 3-20). Norwood, New Jersey: Ablex publishing corporation.

Dörner, D., & Schaub, H. (1994). Errors in planning and decision-making and the nature of human information processing. Applied psychology: an international review, 43(4), 433-453.

Driskell, J. E., & Johnston, J. H. (2006). Stress exposure training. In J. A. Cannon-Browers & E. Salas (Eds.), Making decisions under stress (Vol. 3). Washington, DC: American Psychological Association.

Gohm, C. L., Baumann, M. R., & Sniezek, J. A. (2001). Peronality in extreme situations: thinking (or not) under acute stress. Journal of research in personality, 35, 388-399.

Gonzalez, C. (2005). Decision support for real-time, dynamic decision-making tasks. Organizational Behavior and Human Decision Processes, 96, 142–154.

Gordon, S. E. (1988). Focusing on the human factor in future expert systems. Paper presented at the Artificial Intelligence and other innovative computer applications in the nuclear industry, Snowbird, Utah.

Grootjen, M., Bierman, E. P. B., & Neerincx, M. A. (2006). Optimizing cognitive task load in naval ship control centres: Design of an adaptive interface. Paper presented at the IEA: 16th World Congress on Ergonomics.

Hartanto, D., Kampmann, I. L., Morina, N., Emmelkamp, P. G., Neerincx, M. A., & Brinkman, W. P. (2014). Controlling Social Stress in Virtual Reality Environments. Plos one, 9(3), e92804.

Horsky, J., Schiff, G. D., Johnston, D., Mercincavage, L., Bell, D., & Middleton, B. (2012). Interface design principles for usable decision support: A targeted review of best practices for clinical prescribing interventions. Journal of Biomedical Informatics, 45(6), 1202-1216.

Kenealy, P. M. (1997). Mood state-dependent retrieval: The effects of induced mood on memory reconsidered. The Quarterly Journal of Experimental Psychology: Section A, 50(2), 290-317.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational psychologist, 41(2), 75-86.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. Paper presented at the Proceedings of the Human Factors and Ergonomics Society annual meeting.

Kontogiannis, T., & Kossiavelou, Z. (1999). Stress and team performance: principles and challenges for intelligent decision aids. Safety science, 33, 103-128.

Lerch, F. J., & Harter, D. E. (2001). Cognitive support for real-time dynamic decision making. Information Systems Research, 12(1), 63-82.

McClernon, C. K., McCauley, M. E., O'Connor, P. E., & Warm, J. S. (2010). Stress Training Enhances Pilot Performance During a Stressful Flying Task. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Neerincx, M. A. (2003). Cognitive task load design: model, methods and examples. In E. Hollnagel (Ed.), Handbook of Cognitive Task Design (pp. 283-305). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Peeters, M., Van Den Bosch, K., Meyer, J.-J. C., & Neerincx, M. A. (2014). The design and effect of automated directions during scenario-based training. Computers & Education, 70, 173-183.

Prinsloo, G. E., Derman, W. E., Lambert, M. I., & Rauch, H. G. L. (2013). The effect of a single session of short duration biofeedback-induced deep breathing on measures of heart rate variability during laboratory-induced cognitive stress: a pilot study. Applied psychophysiology and biofeedback 38, 81-90.

Raaijmakers, S. F., Steel, F. W., Goede, M. d., Wouwe, N. C. v., Erp, J. B. F. v., & Brouwer, A.-M. (2013). Heart rate variability and skin conductance biofeedback: A triple-blind randomized controlled study. Paper presented at the Humaine Association Conference on Affective Computing and Intelligent Interaction.

Reason, J. (1987). Cognitive aids in process environments: prostheses or tools? International journal of man-machine studies, 27, 463-470.

Sasou, K., & Reason, J. (1999). Team errors: definition and taxonomy. Reliability engineering and system safety, 65, 1-9.

Schreuder, E. J. A., & Mioch, T. (2011). The effect of time pressure and task completion on the occurrence of cognitive lockup. Paper presented at the Workshop Human Centered Processes.

Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153-189.

Wickens, C. D., Lee, J., Liu, Y., & Becker, S. G. (2004). An introduction to human factors engineering (second ed.). New Jersey: Pearson Education.

8. General Discussion

This final chapter brings together the findings of the previous chapters and, in light of these findings, draws conclusions about the main hypothesis and the research questions. This chapter also reflects on the contributions of this work for the scientific field and, for both trainers and professionals. Furthermore, the overall limitations of the studies are discussed and recommendations on how further research and explorations can continue to improve the COPE feedback system are offered. The chapter ends with a take-home message based on the work presented in this dissertation.

8.1. Conclusions

The experiments conducted for this doctoral dissertation were designed to investigate the following main hypothesis: *"a real-time feedback system improves trainees' performances while working in stressful environments".* As stated in the introduction (Chapter 1) of this thesis, the envisioned feedback system combines biofeedback and more in-depth performance feedback (cf., Kontogiannis and Kossiavelou (1999);Dörner and Schaub (1994)). To analyse the main hypothesis, a model that describes how stress-related cognitive and affective processes influence performances was derived from literature. This model provides the foundation of the feedback system, entailing four main research questions that were studied empirically.

The first research question investigated which aspects of the work content influence the cognitive and affective factors of cognitive task performance. Chapter three shows that cognitive and affective factors are influenced by work content factors such as the task goals and task demands. During a simulated Mars mission that lasted 520 days, six crewmembers divided over two teams performed several tasks every other week. While performing these tasks, their perceived task demand and emotional state were measured. The work content (i.e., mission phase goals and task demands) proved to influence these cognitive and affective factors, and these factors, on their turn, could explain part of the task performance. For example, cognitive task load was higher for a more realistic teaching task than for the less realistic Lunar Lander and Colored Trails games. Arousal levels decreased between the second and third mission phases, suggesting adaptation, or maybe boredom, after the main goal of the Mars mission had been executed. Furthermore, one measure of valence had a positive effect on the performance score of one of the tasks.

The second research question, *Can work content, cognitive and affective factors, measured in real-time, predict trainees' performances in real-time?,* was investigated in Chapters 3 and 4. Both studies found significant predictive models, answering the research question in the affirmative. Statistical results for the simulated Mars mission in

Chapter 3 were meagre; only one measure of valence was considered a predictor of task performance after a Bonferroni correction. Chapter 4 describes an experiment in which two teams of Naval students perform a stressful scenario on two high-fidelity bridge simulators at the Royal Netherlands Naval College. The COPE variables were measured for the different tasks that were performed during this scenario, such as 'navigating the ship through the dark' and 'executing a search-and-rescue operation'. Task performances were rated by a performance score given by the trainers, and by the number of errors made regarding the planning, communication, speed, and task allocation. Trainees' performances rated by trainers were successfully predicted with the COPE variables challenge, threat, and task demand using a linear regression model. The errors made by trainees were predicted with the COPE variables challenge, threat, task demand, perceived task demand, heart rate, and heart rate variability, using a logistic regression model. This study also provided insight into the accuracy of such a prediction. The predictions for performance rated by experts correlated with the observed data ($r = 0.77$) and the error prediction had an accuracy ranging from 66.5% to 91.5%. Although far from perfect, these predictions might be improved by increasing the rate in which the COPE variables were measured.

A feedback system was created based on the COPE model and the predictive models. The COPE FeedBack (COPE-FB) system had three different feedback types: physiological feedback, performance prediction feedback, and error-chance prediction feedback. The design and the use of the COPE-FB system were explained in Chapter 5. The system was evaluated in Chapters 6 and 7.

In Chapter 6 the overall effectiveness of the feedback was tested in the same high-end simulator as was used in the fourth chapter. Trainees received feedback from the COPE-FB system during one half of a scenario. The findings of this study answered the research question *Does providing real-time predictive feedback during stressful events improve trainees' task performance?* Results show that this question can be answered in the affirmative when it comes to reducing the number of errors made regarding planning and speed of task execution. This is also supported by the findings of the laboratory study in Chapter 7. People performed better when they received immediate feedback, than when they did not receive any feedback. These findings can be regarded as an essential first step to improve performance. However, follow up research is needed to determine whether the positive effect found in training with immediate feedback also lasts in the long term. This can be done, for instance, in a follow up training without immediate feedback.

The laboratory study in Chapter 7 also provides a more detailed evaluation of the effect of the individual types and combinations of feedback. It addresses the question: *What type of real-time feedback improves task performances in stressful scenarios?* The analyses in this chapter show that receiving feedback overall improves performances. Effects of specific feedback types were, however, not found on performances, but were found on the usability ratings of the feedback. Participants rated the usability of the physiological feedback on the System Usability Scale (SUS). The SUS score for

physiological feedback was higher compared to situations where physiological feedback was not provided. This SUS rating went down when predicted error-chance feedback was added to the physiological feedback. Individuals indicated that they understood the physiological feedback, explaining the high usability ratings.

The answers to the four research questions led to the acceptance of the main hypothesis. Results show that trainees' performances in a stressful virtual environment indeed improve when they receive real-time feedback. The number of errors decreased (Chapter 6), while the performance scores increased (Chapter 7). A significantly higher usability rating for physiological feedback was found over not receiving physiological feedback, indicating that users understood and preferred this type of feedback.

8.2. Scientific Contribution

The findings described in this doctoral dissertation contribute to different domains of the scientific literature. Human Factors literature already shows theories and models describing the effects of stress on performances, e.g. (Hart & Staveland, 1988; Salas et al., 1996). One contribution of this thesis is that it combines theories and models (Endler & Parker, 1994; Forgas, 1995; Gaillard, 2008; Lazarus, 1999; Mehrabian, 1996; Mosier & Fischer, 2010; Neerincx, 2003; Veltman & Jansen, 2003, 2004b) into one simple model (COPE) that combines information-processing and energetical variables (Robert & Hockey, 1997; Sanders, 1983). Based on the COPE-model, a new feedback system was created that combines biofeedback and predicted performance feedback. Model-based decision support tools are a new development and could therefore benefit from more empirical research (Shim et al., 2002). So, in general, the scientific contribution entails a model, a feedback system, and the empirical findings. The "empirical contributions" can be briefly summarized as follows:

First, the COPE model showed that work content, defined by task goals and task demands, affects cognitive and affective variables. These variables, on their turn, are predictors for trainees' performances. These findings appeared in a unique long duration experiment with extreme demands for the participants.

Second, rich sets of data were collected to make computational models that provide real-time predictions of performance (as rated by experts) and specific errors in stressful simulated environments. Where most descriptive stress models lack a translation into a computational model, most computational models lack the empirical foundation based on experimental data, such as human state and performance data, in either real of virtual environments. The COPE model has both.

Third, the evaluations provide new insights into the effectiveness of real-time feedback. Providing trainees with a combination of predicted performance, predicted error-chance, and physiological real-time feedback improves overall performances, as found in Chapters 6 and 7. Contrary to existing literature (Gonzalez, 2005; Lerch & Harter, 2001), effects for the separate types of feedback were only found for subjective

usability ratings, but not for performance or errors. This shows that subjective preferences do not have to relate to the actual effects on performances.

8.3. CONTRIBUTION FOR TRAINERS

The results of the experiments and the gathered knowledge are not only relevant for other scientists; trainers can also benefit from the results. In Chapter 4, a task analysis resulted in the categorization of errors made during a realistic scenario in a training simulator. The errors made fell into one of the five categories: planning errors, communication errors, task allocation errors, errors regarding speed of task execution, and the remaining (or other) errors. Knowing that most errors fit into one of these categories can help focus training to reduce their occurrence.

The finding that shows how different aspects of work content influence cognitive task load and emotional state of crewmembers differently, is interesting for trainers that help to prepare participants for isolated long-term missions to prepare. Different phases of such missions and the specific tasks that are performed during such missions influence cognitive task load and emotional state. The trainees should keep in mind that task goals are related to the training goal and how goals change over time.

Furthermore, the introduction of feedback systems into virtual environments will reduce the trainer's workload. Trainees can use feedback to recognize and identify their pitfalls when they are under stress, leaving more room for the trainers to focus on other learning aspects of the training. Future research should investigate this opportunity in more detail.

8.4. LIMITATIONS

One limitation, concerning the accuracy of the predictive models, is relevant to the findings of Chapters 4 and 6. The performance predictions are based on subjective performance ratings and subjective comments of trainers. First of all, variation between trainers on how they perceive a performance affects the reliability of this measure. Next, errors are derived from the trainers' comments and are therefore based on what they consider an error and what they anticipated would eventually result in negative consequences. Therefore the model's predictions are limited to the accuracy of trainers' interpretation of the situation. Automatic measures for performances could increase the consistency of what is considered an error and what is considered a specific rating of performance. In Chapter 7, errors were automatically recognized after the experimenters determined what was considered to be a specific error. This method increased the consistency but still based performance levels on subjective assessments of performance.

Furthermore, the studies in this dissertation are limited to experimental training settings. The current work does not provide empirical findings regarding the transfer of

skills learned in a simulator to settings outside the simulators. The experiments in Chapter 6 and Chapter 7, both showed a performance improvement while using ship simulators differing in fidelity and realism. When presenting the COPE-FB system in a more realistic, high-fidelity simulator such as in Chapter 6, the results showed a decrease in planning errors and errors regarding the speed of task execution, whereas the less realistic, low-fidelity simulator from Chapter 7 showed an overall performance score increase.

Since responses to virtual environments are more realistic when the environment is more realistic (Slater, Khanna, Mortensen, & Yu, 2009) the results from Chapter 6 are probably most reliable in their transfer to a real ship setting. Previous research by Neerincx, Kennedie, Grootjen, and Grootjen (2009) showed that a model that predicted performance was able to predict performances both in a ship simulator and on a real ship environment, with some decrease in model fit in the real world. This supports the hypotheses that the results from Chapter 6 are likely to transfer to a real world setting.

The predictive models are based on groups containing trainees with the same level of experiences for the task at hand. Although no specific investigations were made regarding the COPE model's adaptability to different experienced professionals, literature shows that professionals greatly rely on their experiences when making decisions under stress (Klein, 1993; Noble, 1998). This indicates that not only task characteristics affect the COPE variables, but different levels of experience have an effect as well. Further research is necessary to investigate the relevance and manifestation of this effect, and to test to what extent the predictive models generalize the different levels of experience the professionals have.

Chapter 2 explains the COPE model, including the variable 'coping strategy'. This variable was not included in the studies in this thesis, since the questionnaires that measure coping strategy were inappropriate and obtrusive in the current settings. While it was argued that coping strategy might be predicted with individual characteristics, Cohen and Lazarus (1973) and Folkman and Lazarus (1985) state that coping behaviour is a dynamic process that cannot be predicted by looking at personality characteristics. Delahaij (2009), on the other hand, found that individuals who had a specific coping style before the training mainly use this coping style during the training in which they endure acute stress. This suggests that coping style (general style in which individuals work and cope) can be measured beforehand and can thus be included in the COPE-FB system. For example, the models can be calibrated for trainees with a more emotion-focused copying style or with a more task-focused coping style. The model that best suits the coping style of a specific trainee can then be selected in the COPE-FB system.

8.5. FUTURE OF THE COPE-FB SYSTEM

The work in this thesis provides suggestions for improvements on both the COPE model and the COPE-FB system. Task characteristics (as mentioned in Section 8.4), level of

experience, and, probably, coping style could have an impact on the cognitive and affective variable. In this thesis, the COPE model functions were adapted for different scenarios and training settings, but might benefit from more specific calibration for different tasks, levels of experience, and personal coping style as well. The predictive models themselves are currently comprised of regression models, while other models such as Bayesian networks are also able to predict performance out of cognitive task load (Neerincx et al., 2009a). Chapters 4 and 7 both illustrate how the predictive models were established and how accurate they are. The accuracy varies between the two chapters and might be increased by basing the models on a larger sample size. Another way to increase the models' predictive accuracy might be to add extra physiological variables, which might also increase the objectiveness of the arousal measurements.

Improvement suggestions emerging from Chapters 6 and 7 concern the design and use of the COPE-FB system. Both chapters state that a more detailed explanation is necessary for the users of the system to fully understand the different types of feedback and what they entail. Regarding the use of the system, implementing it into a mobile or hand-held device will increase the accessibility to the feedback to trainees who are not always behind a fixed computer screen. Another option is to add a warning signal to indicate when performance levels have reached a certain threshold. Adding a warning signal to a hand-held device such as a tactical signal will call a trainee to attention without interrupting other trainees.

Before the COPE-FB system is ready to be used outside of a simulator, other modifications are also necessary. The values of the COPE variables appraisal and task demand were determined beforehand during training, and were applied in the COPE-FB system during training. Further research is necessary to ascertain to what extent predetermined values of appraisal and task demand cover these values in real-life settings. One way to bypass these predetermined values could be by using automatic or objective measures for appraisal and task demand (Grootjen, Neerincx, van Weert, & Truong, 2007). Such measures will make the system ready and usable for evaluations in real world settings.

8.6. TAKE HOME MESSAGE

The COPE model shows mechanisms that explain how stress factors influence performance, leading to the design and creation of the COPE FB system. When trainees receive feedback through this system, either in a high-end virtual environment or in a more controlled low-end virtual task, they make less errors regarding planning and speed of task execution and their overall performance ratings increased. Also, trainees indicated that they preferred receiving physiological feedback over not receiving physiological feedback. This is also reflected in the higher subjective usability ratings for this type of feedback.

In the introduction of this thesis, the 1988 incident with the USS Vincennes (figure 1.1) was described. While in combat, this Naval warship misidentified an

approaching commercial airliner for an attacking F-14 fighter aircraft and shot it down, killing 290 civilians and crewmembers. Of course, the COPE model does not include all the factors that played a role during this incident. However, results of the studies in this thesis do show that COPE-based feedback can improve team members' performances while working in similar virtual situations with high cognitive load. If the COPE-FB system continues to be improved and is added to training sessions in virtual environments, it is expected that in the near future professionals working in stressful, dangerous settings can receive improved and extensive virtual preparation for real-life stressful scenarios.

REFERENCES

Cohen, F., & Lazarus, R. S. (1973). Active coping processes, coping dispositions, and recovery from surgery. *Psychosomatic Medicine, 35*(5), 375-389.

Delahaij, R. (2009). *Coping under acute stress: the role of person characteristics.* Kon. Broese & Peereboom, Breda.

Dörner, D., & Schaub, H. (1994). Errors in planning and decision-making and the nature of human information processing. *Applied psychology: an international review, 43*(4), 433-453.

Endler, N. S., & Parker, J. D. (1994). Assessment of multidimensional coping: Task, emotion, and avoidance strategies. *Psychological assessment, 6*(1), 50.

Folkman, S., & Lazarus, R. S. (1985). If it changes it must be a process: study of emotion and coping during three stages of a college examination. *Journal of personality and social psychology, 48*(1), 150.

Forgas, J. P. (1995). Mood and judgement: the affect infusion model (AIM). *Psychological bulletin, 117*(1), 39-66.

Gaillard, A. W. (2008). Concentration, stress and performance. *Performance under stress*, 59-75.

Gonzalez, C. (2005). Decision support for real-time, dynamic decision-making tasks. *Organizational Behavior and Human Decision Processes, 96*, 142–154.

Grootjen, M., Neerincx, M. A., van Weert, J. C., & Truong, K. P. (2007). Measuring cognitive task load on a naval ship: implications of a real world environment *Foundations of Augmented Cognition* (pp. 147-156): Springer.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology, 52*, 139-183.

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making *Decision making in action: Models and Methods*: Ablex Publishing Corporation.

Kontogiannis, T., & Kossiavelou, Z. (1999). Stress and team performance: principles and challenges for intelligent decision aids. *Safety science, 33*, 103-128.

Lazarus, R. S. (1999). *Stress and emotion: a new synthesis*. New York: Springer Publishing Company, Inc.

Lerch, F. J., & Harter, D. E. (2001). Cognitive support for real-time dynamic decision making. *Information Systems Research, 12*(1), 63-82.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology, 14*(4), 261-292.

Mosier, K. L., & Fischer, U. (2010). The role of affect in naturalistic decision making. *Journal of cognitive engineering and decision making, 4*(3), 240-255.

Neerincx, M., Kennedie, S., Grootjen, F., & Grootjen, M. (2009). *Modelling the Cognitive Task Load and Performance of Naval Operators.* Paper presented at the Lecture Notes in Artificial Intelligence. Schmorrow, DD; Estabrooke, IV; Grootjen, M.(Eds.), Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience. Proceedings of the 5th International Conference of the Augmented Cognition.

Neerincx, M. A. (2003). Cognitive task load design: model, methods and examples. In E. Hollnagel (Ed.), *Handbook of Cognitive Task Design* (pp. 283-305). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Noble, D. (1998). *Distributed situation assessment.* Paper presented at the Proc. FUSION.

Robert, G., & Hockey, J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology, 45*(1), 73-93.

Salas, E., Driskell, J. E., & Hughes, S. (1996). Introduction: the study of stress and human performance. In J. E. Driskell & E. Salas (Eds.), *Stress and Human Performance* (pp. 1-45). Hillsdale, NJ: Erlbaum.

Sanders, A. (1983). Towards a model of stress and human performance. *Acta psychologica, 53*(1), 61-97.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present and future of decision support technology. *Decision Support Systems, 33*, 111-126.

Slater, M., Khanna, P., Mortensen, J., & Yu, I. (2009). Visual realism enhances realistic response in an immersive virtual environment. *Computer Graphics and Applications, IEEE, 29*(3), 76-84.

Veltman, J. A., & Jansen, C. (2003). Differentiation of mental effort measures: consequences for adaptive automation. In G.R.J.Hockey, A. W. K. Gaillard & O. Burov (Eds.), *Operator Functional State: The Assessment and Prediction of Human Performance Degradation in Complex Tasks* (pp. 249-259). Amsterdam: IOS Press.

Veltman, J. A., & Jansen, C. (2004). The adaptive operator. *Human performance, situation awareness, and automation: Current research and trends, 2*, 7-10.

SUMMARY

Professionals working in different domains often experience stressful conditions evoked by disasters or crisis scenarios. Regardless of these conditions, they have to perform at high standards in order to preserve safety for themselves, avoid any casualties, and to resolve the overall situation. Stress, however, negatively affects cognitive processes and thereby decreases performances. This doctoral thesis aims to improve professionals' decisions and performances when working in risk- and stressful situations. To reach this goal, a model-based support system was constructed and evaluated.

Based on existing models and theories, a model was created to portray the process of performing under stress. The COgnitive Performance and Error (COPE) model explains how work content influences cognitive and affective factors and how, in turn, these factors affect task performances. This model was evaluated during two experiments. First, in a long-term simulated Mars mission, experimental tasks were executed every two weeks. Before, during, and after the tasks, participants subjectively reported their cognitive and affective measures. Results showed that COPE variables differed when work content varied, indicating that work content indeed influenced cognitive and affective variables. Results also showed that task performance could be explained by some cognitive and affective variables. Next, the COPE model was fitted on data collected during a stressful scenario in a high-fidelity Naval simulator. This experiment showed that during virtual training, the COPE-variables were predictors for performances. This resulted in models that could predict performance scores and the number of errors made.

The performance predicting models established in the scenario-based Naval training were implemented into a feedback system. This COPE-based FeedBack system (COPE-FB system) provided physiological feedback (heart rate measure), performance prediction feedback, and feedback on the predicted error chances in real-time. When trainees in the high-fidelity Naval simulator received feedback from the COPE-FB system, the amount of errors regarding the planning and speed of task execution decreased. In a laboratory study, participants were confronted with different combinations of the three feedback types while participating in a stressful fire extinguish simulation. Although performances did not improve when the effect of separate feedback types were analysed, all feedback combinations as a whole resulted in an increased performance score. Another result from this study illustrated that participants preferred to receive only the physiological feedback.

The studies in this thesis show that the COPE model can be translated into predictive models that use real-time variables to predict performances. Implementing such models into a feedback system resulted in a feedback system that decreased errors in a scenario-based Naval simulator training. In a low-fidelity laboratory study, all feedback combinations in one factor increased overall performance scores. This thesis shows that the COPE-FB system increases parts of trainees' performances in stressful virtual environments. It also gives some suggestions on how the system can be improved to further increase the trainees' performances.

SAMENVATTING

Professionals in verschillende velden krijgen vaak te maken met stressvolle situaties veroorzaakt door crises. Onder zulke omstandigheden moeten ze goed presteren om de crises op te lossen en veiligheid te creëren voor zichzelf en omstanders. Helaas heeft stress vaak negatieve gevolgen op cognitieve processen waardoor prestaties dalen. Deze thesis probeert de prestaties van professionals die werken onder stress weer te verbeteren met behulp van een modelgebaseerd systeem. De ontwikkeling en evaluatie worden aan de hand van onderzoeken uitgelegd.

Een model was gecreëerd, gebaseerd op bestaande modellen en theorieën. Dit model laat het proces van presteren onder stress zien. De "Cognitieve Prestatie en Fouten" (COPE) model legt uit dat werkinhoud invloed heeft op cognitieve en affectieve factoren en dat deze factoren weer invloed hebben op taak prestaties. Dit model is vervolgens geëvalueerd aan de hand van twee experimenten. Tijdens een gesimuleerde Mars-missie werden tweewekelijks taken uitgevoerd. Voor, tijdens en na de taken werden cognitieve en affectieve factoren subjectief gemeten. Per taak waren er ook prestatiescores berekend. Resultaten van dit onderzoek lieten zien dat de werkinhoud factoren inderdaad invloed hebben op de cognitieve en affectieve factoren van het COPE model. In de resultaten was ook te zien dat de prestatiescores voorspeld konden worden vanuit de cognitieve en affectieve factoren. Vervolgens is het model getest tijdens een virtuele training bij de Marine. Marine-studenten voerden een stressvolle taak uit in de brugsimulator en gaven aan wat de level van de COPE-variabelen was tijdens het uitvoeren van deze taken. Data van dit onderzoek resulteerde in verschillende voorspellende modellen. Modellen konden de prestatiemaat en het aantal gemaakte fouten voorspellen.

Deze voorspellende modellen werden vervolgens geïmplementeerd in een feedback systeem. Dit COPE-gebaseerde feedback systeem (COPE-FB systeem) geeft fysiologische feedback (hartslagweergave), voorspelde prestatie-feedback, en voorspelde kans-op-fouten feedback, allen in real time. In de highfidelity Marine simulator werden studenten tijdens een nieuw experiment geconfronteerd met dit feedback system. Uit de resultaten bleek dat deze studenten minder fouten maakten als zij feedback ontvingen over hun prestaties. Het COPE-FB systeem werd nogmaals getest in een laboratorium onderzoek waarbij proefpersonen een stressvolle brandblus-taak uitvoerden. Zij werden geconfronteerd met de losse feedback-typen en verschillende combinaties van twee typen feedback. Prestatiescores verbeterden als alle feedback samen werd genomen in analyses, maar de verschillende feedback combinaties hadden alleen invloed op de usability-maten. Proefpersonen gaven aan dat ze de fysiologische feedback het fijnst vonden.

De onderzoeken beschreven in deze thesis lieten zien dat het COPE model in staat is om prestaties te voorspellen in real time. Het implementeren van dergelijke modellen in een feedback systeem resulteerde in verlaging van het gemaakte aantal fouten in een Marine simulator en een verbetering in prestatiescores tijdens een brandblus-taak. Hoewel er enkele beperkingen waren in de onderzoeken zijn de resultaten veelbelovend. Met de gemaakte suggesties kan het COPE-FB systeem verder worden verbeterd, waardoor het in de toekomst de prestaties van studenten nog meer kan verhogen.

ACKNOWLEDGEMENT

Na het afronden van mijn afstudeer thesis, hét sociale robot onderzoek, nodigde Prof. dr. Mark A. Neerincx mij uit om te solliciteren op een promotie project waar hij samen met Dr. Ir. Willem-Paul Brinkman het voorstel voor had geschreven. Aangezien mijn afstudeer onderzoek zo soepel was verlopen en leuk was om uit te voeren leek het me gelijk al een goed idee om verder te gaan in de academische wereld. Het promotie onderzoek bleek al snel een stuk gecompliceerder maar samen met de dagelijkse (wekelijkse) begeleiding van Willem-Paul en Mark heb ik mooie resultaten bemachtigd, interessante onderzoek locaties bezocht (ESA, het KIM) en deze dissertatie kunnen schrijven. Bedankt Mark en Willem-Paul, voor de begeleiding, sturing en nodige discussies.

I would like to thank all of the professors in my committee: Prof. Dr.-Ing. F. Flemisch and Prof. dr. ir. S.A. Meijer, thank you for joining my defence in The Netherlands. Prof. dr. J.M.C. Schraagen van TU Twente, Prof. dr. Ir. P. A. Wieringa en Prof. dr. C. M. Jonker, beide van TU Delft. I appreciate the time you all took to read my dissertation and place comments where necessary. I am looking forward to the questions and discussions during the defence.

Het project beschreven in deze dissertatie maakte deel uit van het Nationaal Initiatief Hersenen en Cognitie, onder de pijler Veiligheid. Ook al zag ik ze niet vaak, het was altijd fijn om positieve feedback te ontvangen van prof. dr. F. Leeuw en Kathy de Kogel tijdens voortgang presentaties.

Twee van de vier onderzoeken beschreven in mijn dissertatie waren niet mogelijk geweest zonder de gastvrijheid en behulpzaamheid vanuit het KIM oftewel, the Royal Netherlands Naval College. Heel veel dank voor het openstaan voor mijn onderzoek en de hulp ter plaatse; docent brugsimulator (J.A.) Rico Bloemberg, luitenant-ter-zee 2OC J. (Job) van Rooijen, Ronald Kempenaar, Rick, en natuurlijk alle andere trainers van die tijdens mijn experimenten in 2013 en in 2014 hebben geholpen.

De afgelopen vier jaar was ik voornamelijk te vinden bij TNO. Hier heb ik samengewerkt aan interessante projecten met Victor Kallen, Nelleke van Wouwe, Peter-Paul van Maanen, Nanja Smets en Jurriaan van Diggelen. Ook heb ik van Marc Grootjen veel adviezen en tips gekregen wat betreft het gebruik van de Mobi8 apparaten. Ook wil ik graag Arnold Dittmar, Bart Vastenhouw & Ruud de Jong bedanken voor het uitlenen van talloze displays, laptoppen en de bijbehorende kabels.

Een aantal studenten hebben op verschillende manieren geholpen bij mijn onderzoeken. Nadia den Braber, heel erg bedankt voor je hulp bij de Mars500 analyses. Ik hoop dat je er iets aan had en dat je de baan hebt gevonden waar je naar op zoek was. Sophie Zuiderduin heeft bij TNO het 'brandjestaak' opnieuw leven in geblazen en veel werk verricht voor hoofdstuk 7. En natuurlijk Dustin Lim. Hij heeft als student-assistent hard gewerkt om het feedback system te laten functioneren zoals wij voor ogen hadden. En dat is gelukt!

During the weekly VRET meetings there were always interesting presentations from: Corine Horsch, Myrthe Thielman, Dylan Schouten, Dwi Hartanto, Alex Kayal, Ni Kang and Wenxin Wang. I hope you will all enjoy your last few steps in your PhD adventures and good luck with all the following steps in your post-PhD-lives.

Corine Horsch verdient nog een extra bedankje voor de gezellige ritjes tussen Delft en Utrecht. Onze carpool ritjes maakte een toch wel lange rit een stuk aangenamer. Hoe bevalt je nieuwe auto? At TNO I very much enjoyed the adventures in the Soesterbergse forests during lunch time with Bruno & Ksander. Hopefully we will meet again soon, whether it is in Madrid or in Tübingen.

Verder wil ik Glenn bedanken voor zijn professionele input in het design van de kaft. Kom snel maar eens langs in Tübingen, dan kunnen we een foto sessie houden in de bergen, want die zijn hier namelijk!

Een betere 'stress-relief' dan het COPE-FB systeem is mogelijk gemaakt door Nienke, Naomi, Glenn, Joris, AJ, Marco en Elja. Door middel van een kopje thee, glaasje wijn, een potje roller derby of een repetitie met de band (later vervangen door gewoon showtjes bezoeken) hebben deze lieve vrienden voor een hoop ontspanning gezorgd.

Mama en papa, bedankt voor jullie steun en toewijding tijdens mijn studie jaren. Ik heb dit altijd zeer op prijs gesteld. En mijn lieve zusje Dana, bedankt voor je hulp met het leesbaar maken van de lange zinnen. Ik kijk uit naar de familie reünie en ik hoop dat je snel een baan vindt waarbij je je taalvaardigheden kan benutten. Bedankt dat je ook nog eens paranimf wilt zijn.

Joris, na de vele optredens in binnen- en buitenland waarbij jij op de voorgrond stond, zijn de rollen nu omgekeerd. Ik ben blij dat je een paranimf wilt zijn en ik hoop dat het pak niet al te verschrikkelijk zit. Gelukkig wordt het vast gelegd op foto, zodat iedereen dit spektakel kan waarnemen.

Behalve een dr. titel heb ik nog iets overgehouden aan het promoveren; Ksander. In jouw dankwoord heb je het over roadtrips, maar eerlijk gezegd ben ik blij dat we niet meer maandelijks 1200 km in ons eentje hoeven af te leggen. Voortaan leggen we ze samen af.